

**Validation of a rating scale
for distance education university student
essays in a literature-based module**

by

Maxine Welland Ward-Cox

Submitted In fulfilment of the requirements for the Degree of

Doctor of Literature and Philosophy

in the subject of

English

at the

University of South Africa (Unisa)

Supervisor: Professor Brenda Spencer

Co-Supervisor: Dr Ruth Scheepers

January 2020

DECLARATION

Student Name: Maxine Welland Ward-Cox

Student Number: 02311194

Degree: PhD (LAN LIN & LIT)

Title: Validation of a rating scale for distance education university student essays in a literature-based module

I declare that the above thesis is my own work and that all the sources that I have used or quoted have been indicated and acknowledged by means of complete references

I further declare that I have not previously submitted this work, or part of it, for examination at Unisa for another qualification or at any other higher education institution

Signed: *M Ward-Cox* Date: *3 January 2020*

ABSTRACT

This thesis reports on the findings of a study to validate an assessment scale for writing in an Open Distance Learning (ODL) context by first-year students in their responses to English literary texts. The study involved the interrogation of an existing scale, adapted from Jacobs *et al.* (1981), which was being used for the *Foundations in English Literary Studies* (ENG1501) module at the University of South Africa. Despite the credibility of the original scale, the modified version had been used in language- and literature-based courses in the English Studies Department since 1998 and had not been updated or empirically tested in the context of the target group. Thus, the gap that this current study addressed was the need for a valid rating scale that takes into account the complexities of literature teaching and ODL in the current South African university environment. This thesis includes a review of the debate on validity and the validation of rating scales both internationally and in South Africa, the ODL environment, and the assessment of assignments based on literary texts, particularly in the multicultural South African context. The methodology included research of both a quantitative and a qualitative nature. The outcome was an empirically-validated scale that should contribute to the quest for accuracy in assessing academic writing and meet the formative and summative assessment needs of the target group.

Key terms: validity, validation, assessment, rating scale, distance education, open distance learning (ODL), open distance e-learning (ODeL), academic writing skills, South African students.

ACKNOWLEDGEMENTS

I wish to acknowledge a large debt of gratitude to:

- Professor Faans Steyn for his expert statistical assistance
- Ms Jackie Viljoen and Mr Desmond Collier for their professional and meticulous reference editing and proofreading of this thesis, and for their advice and encouragement
- Mr Robert Ward-Cox for his unflagging support and practical assistance
- Dr Malcolm Venter, Ms Morag Venter, Ms Lindsey Lewis and Mr Des Collier for their constant support and encouragement throughout the course of this project
- All tutors, markers and panel members, especially Dr Christine Marshall, Dr Malcolm Venter, Dr Diana Mc Dermott, Dr Stephan Maritz, Ms Morag Venter, Ms Lindsey Lewis, Mr Desmond Collier, Ms Ricky Woods, Ms Anne Peltason, Mr Jaques Du Toit, Ms Trunell Morom, Mr Reinhardt Fourie, Ms Devarshinee Chetty, Ms Christin Williams, and Ms Miemie Taljaard
- My grandparents, parents and step-parents who taught me the value of education
- Professor Brenda Spencer and Dr Ruth Scheepers for their excellent supervision of this thesis, their incisive feedback, and their invaluable guidance without which this study would not have been possible.

Letter of Confirmation of Statistical Analyses



NORTH-WEST UNIVERSITY
YUNIBESITHI YA BOKONE-BOPHIRIMA
NOORDWES-UNIVERSITEIT
POTCHEFSTROOM CAMPUS

Private Bag X6001, Potchefstroom
South Africa 2520

Tel: (018) 299-1111/2222

Web: <http://www.nwu.ac.za>

Statistical Consultation Services

Tel: (018) 299-2180

Fax (018) 299 2557

E-Mail: faans.steyn@nwu.ac.za

Re: Ms Maxine Ward-Cox

I hereby confirm that I have assisted Ms Maxine Ward-Cox, student number 02311194, with the statistical aspects of her PhD thesis at the University of South Africa with title: VALIDATION OF A RATING SCALE FOR DISTANCE EDUCATION UNIVERSITY STUDENT ESSAYS IN A LITERATURE-BASED MODULE

Kind regards,

A handwritten signature in black ink, appearing to read 'Faans Steyn'.

Prof Faans Steyn (PhD, Pr. Sci. Nat)

Statistical consultant

CONTENTS

DECLARATION	ii
ABSTRACT.....	iii
ACKNOWLEDGEMENTS	iv
Letter of Confirmation of Statistical Analyses.....	v
CONTENTS	vi
LIST OF TABLES.....	xii
LIST OF FIGURES	xiii
1. CHAPTER 1: INTRODUCTION.....	14
1.1 RESEARCH OVERVIEW.....	14
1.2 BACKGROUND AND RATIONALE	14
1.3 THESIS STATEMENT	17
1.4 STATEMENT OF THE RESEARCH PROBLEM	17
1.5 RESEARCH QUESTIONS.....	18
1.6 AIM AND OBJECTIVES OF THE STUDY.....	19
1.7 NATURE OF THE RESEARCH	19
1.7.1 Theoretical underpinning	20
1.7.2 Research methodology	20
1.8 ETHICAL CONSIDERATIONS.....	24
1.9 OUTLINE OF CHAPTERS.....	25
1.10 CONCLUSION.....	27
2. CHAPTER 2: CONTEXT OF THE RESEARCH – TARGET GROUP AND MODULE	29
2.1 INTRODUCTION.....	29
2.2 DISTANCE EDUCATION AND LANGUAGE LEARNING.....	30
2.3 OPEN DISTANCE LEARNING (ODL) AND OPEN DISTANCE E-LEARNING (ODEL).....	35
2.4 ASSESSMENT AND FEEDBACK IN DISTANCE LEARNING	38
2.5 THE RATING SCALE AS FEEDBACK IN FORMATIVE ASSESSMENT	45
2.6 Target module	50
2.7 READING AND WRITING SKILLS IN THE CURRENT SITUATION.....	52
2.7.1 Reading skills	53
2.7.2 Academic writing in theory and practice	55
2.8 LITERARY STUDIES AND ASSESSMENT	58
2.9 CONCLUSION.....	63

3.	CHAPTER 3: THEORIES OF VALIDITY AND RELIABILITY: CHANGING PERSPECTIVES	64
3.1	INTRODUCTION	64
3.2	THEORIES OF VALIDITY: CHANGING PERSPECTIVES.....	64
3.2.1	Definitions of validity	64
3.2.2	Types of validity	67
3.3	MOVING TOWARDS A UNIFIED INTERPRETATION OF VALIDITY	74
3.3.1	Messick's unitary approach to validity	75
3.3.2	Critique of Messick's theory	77
3.4	VALIDITY AND RELIABILITY	88
3.5	CONCLUSION.....	92
4.	CHAPTER 4: THE VALIDATION PROCESS.....	94
4.1	INTRODUCTION	94
4.2	DEFINITIONS OF VALIDATION	94
4.3	THE VALIDATION PROCESS	96
4.3.1	The argument - based approach to validation	100
4.3.2	Challenges relating to determining the construct for ENG1501.	104
4.4	MODELS AND FRAMEWORKS	108
4.4.1	Bachman's model of communicative ability	108
4.4.2	Bachman and Palmer	111
4.4.3	Cambridge ESOL framework	112
4.4.4	Weir's framework (2005)	113
4.4.5	The framework of Shaw and Weir (2007)	115
4.5	CRITERIA AND BAND SCALES.....	117
4.5.1	Level descriptors and band levels	118
4.5.2	Band levels	121
4.6	FACTORS THAT IMPACT UPON SCORES.....	124
4.6.1	Factors directly related to learning, teaching and assessment	125
4.6.2	Administrative setting	130
4.6.3	Impact on society	133
4.7	THE RATING SCALE 'MYTH': FURTHER OBSERVATIONS ON RATING SCALES.....	134
4.8	CONCLUSION.....	138
5.	Chapter 5: Research Methodology	139
5.1	Introduction.....	139
5.2	RESEARCH QUESTIONS.....	139
5.3	RATIONALE: RESEARCH DESIGN AND METHODS	140

5.4	DATA	145
5.5	SAMPLING METHOD.....	149
5.6	SAMPLE POPULATION	150
5.6.1	Sample composition by home language	151
5.6.2	Sample composition by province and residential regional office	152
5.6.2	Sample composition by race and gender	153
5.7	MEASURING INSTRUMENTS	154
5.8	RESEARCH PROCEDURE	160
5.8.1	Pilot study	162
5.8.2	Main study	164
5.8.3	The design stage	172
5.8.4	The construction stage	173
5.8.5	Testing of the revised scale	173
5.9	ETHICAL CONSIDERATIONS.....	174
5.9.1	Consent and voluntary participation	174
5.9.2	No harm to participants	174
5.9.3	Confidentiality	174
5.10	OVERVIEW: RELEVANCE OF METHODS TO RESEARCH QUESTIONS	175
5.11	SUMMARY OF PROCESS	177
5.12	CONCLUSION.....	178
6.	CHAPTER 6: RESEARCH FINDINGS: EXISTING SCALE.....	179
6.1	INTRODUCTION	179
6.2	SELECTION AND DOWNLOADING OF SCRIPTS.....	179
6.3	PILOT STUDY	179
6.3.1	Statistical results – analysis of variance (20 scripts)	180
6.3.2	Comments from pilot study markers	181
6.4	MAIN STUDY (60 SCRIPTS, INCLUDING SCRIPTS 1-20)	183
6.4.1	Collect data and mark scripts	183
6.4.2	Analyse data	183
6.4.3	Explanation of the results	183
6.4.4	Comments by markers of the main study	192
6.4.5	Questionnaires	195
6.4.6	Summary of feedback	196
6.5	OBSERVATIONS ON QUESTIONNAIRE FEEDBACK.....	203
6.5.1	Number of levels	203
6.5.2	Weighting of marks between content and language use	204
6.5.3	The extent to which markers adhered to the marking scale	205

6.5.4	The extent to which the scale should take the distance learning context into account.....	205
6.5.5	The extent to which the scale should take the multicultural and multilingualistic target market into account.	205
6.5.6	The perceived subjectivity of the scale	207
6.6	CONCLUSION.....	207
7.	CHAPTER 7: DEVELOPMENT OF THE NEW SCALE.....	210
7.1	INTRODUCTION	210
7.2	Design and construction of the proposed new scales	210
7.2.1	The design of the scale	217
7.3	CONSTRUCTION OF THE REVISED VERSION OF EXISTING SCALE ...	229
7.4	TESTING OF THE REVISED SCALE (MODEL 1)	231
7.4.1	Marking of scripts	231
7.4.2	Quantitative analysis	231
7.4.3	Qualitative findings	233
7.4.4	Final questionnaires – feedback on Model 1.	235
7.4.5	Issues raised by the answers to the questionnaire (Model 1)	238
7.5	CONSTRUCTION OF THE TWO-DIMENSIONAL GRID	240
	Proposed Marking Grid ENG1501 Literary Assignments Version 1.....	241
7.6	TESTING THE TWO-DIMENSIONAL GRID	242
7.6.1	Marking	242
7.6.2	Quantitative testing	242
7.6.3	Qualitative findings: Markers' comments (Model 2)	244
7.7	Summary of results of questionnaire – Model 2	245
7.7.1	Issues arising from responses to the questionnaire	248
7.8	ACTIONS AND SUGGESTIONS ARISING FROM FEEDBACK MARKERS AND PANEL MEMBERS.....	250
7.8.1	Assessing the essay-type question	250
7.8.2	The use of graphics for formative assessment	251
7.9	CHOICE OF RATING SCALE.....	253
7.10	SUGGESTED IMPLEMENTATION.....	253
7.11	CONCLUSION.....	263
8.	Chapter 8: Conclusion and recommendations.....	264
8.1	INTRODUCTION	264
8.2	QUESTIONS	265
8.2.1	Sub- question 1	266
8.2.3	Sub- question 3	268
8.2.4	Sub- question 4	270

8.2.5 Sub- question 5	271
8.3 RECOMMENDATIONS.....	272
8.3.1 The validity of the existing scale	272
8.3.2 The complex ODL, multilingual and multicultural environment	273
8.3.3 Number of levels	274
8.3.4 Weighting of marks	274
8.3.5 Plagiarism	275
8.3.6 Subjectivity	275
8.3.7 Formative assessment	275
8.3.8 Type of scale	276
8.4 REPONSES TO THE PRIMARY RESEARCH QUESTIONS	277
8.5 LIMITATIONS OF THE RESEARCH.....	277
8.6 RECOMMENDATIONS FOR FURTHER STUDY	279
8.6.1 Research on student input	279
8.6.2 Use of different genres and target groups	279
8.7 CONCLUSION.....	279
List of References.....	281
Appendix A: Participation and Informed Consent Leaflet	305
Appendix A1: Research Permission.....	308
Appendix A2: Participant's Consent Form.....	309
Appendix A3: Deelnemer se Toestemmingsvorm	310
Appendix A4: Ifomu Yemvume Yomthabathi Nxaxheba.....	311
Appendix B: Marking Grid ENG1501	312
Appendix C: MODULE FORM	314
Appendix D1: Email to Students	316
Appendix D2: Email to Unisa Markers and Tutors.....	317
Appendix E: Assignment Memorandum	318
Appendix F: Rasch Analysis: Existing Rating Scale.....	321
Appendix G: Extracts from correspondence with panel members	325
Appendix H : Statistical information: Model 1	327
Appendix I : Statistical Results of Model 2	329
Appendix J: Final Grid Model 1	331
Appendix K : Proposed Marking Grid ENG1501 Literary Assignments	332
Appendix L Statement of of originalty of topic.....	333
Appendix M Personal Details	334
Appendix N: Turnitin similarity index showing single source similarities (5% and above) excluding referenced quotations.....	337

LIST OF TABLES

Table 3.1: Facets of test validity.....	75
Table 3.2: Interpretation of Messick's validity matrix	79
Table 3.3: The relationship of a selection of fundamental considerations in language testing.....	81
Table 4.1: Common European framework (Table vi) – global scale (Council of Europe, 2001)	119
Table 4.2: IELTS writing band level descriptors for bands 8 and 9.....	120
Table 4.3: Summary of criteria distinguished in four current rating scales	121
Table 4.4: IELTS band scale descriptors.....	122
Table 4.5: The Jacobs' scoring profile.....	123
Table 4.6: Language Use Level 3 of the scale used for the target module	124
Table 5.1: Module Assignment.....	146
Table 5.2: Distribution of sample scripts according to levels of the existing rating scale	149
Table 5.3: Sample composition by language for Semester 1 and 2, 2016 (as at 27 October 2016).....	151
Table 5.4: Sample composition by province and residential regional office for Semester 1 and 2, 2016 (as at 27 October 2016).....	152
Table 5.5: Sample composition by race and gender for Semester 1 and 2, 2016 (as at 27 October 2016	154
Table 5.6: Markers' questionnaire	156
Table 5.7: Questionnaire for feedback on Models 1 and 2	158
Table 5.8: Marking grid for ENG1501.....	161
Table 5.9: Example reliability index for data new scale trial Model 2	171
Table 6.1: Script and marker reliability: Total marks, Markers 1 – 10 (60 scripts).....	191
Table 6.2: Summary of responses to questionnaire (existing scale).....	196
Table 7.1: Comparative extracts from existing marking grids	213
Table 7.2: Extract showing amended categories.....	218
Table 7.3: Final draft of proposed Likert Scale assessment grid	219
Table 7.4: Extract from revised grid (Model 1).....	223
Table 7.5: De Beers English Olympiad Marking Grid.....	225
Table 7.6: Marking Guide: Department of English, Nelson Mandela University	227
Table 7.7: Revised scale: Model 1 (Draft 1) Total 50 marks	230

Table 7.8: Reliability Model 1	233
Table 7.9: Summary of responses to questionnaire: Model 1	235
Table 7.10: Model 1 grid: Final version 1	239
Table 7.11: Model 2 grid	244
Table 7.12: Summary of reliability of Model 2.....	244
Table 7.13 Comparative summary of the quantitative results of Models 1 and 2.....	244
Table 7.14 Summary of results of questionnaires.....	245
Table 7.15: Results of assessing essay-type answers	251
Table 7.16: Grid for assessing Grade 12 literary essays	252
Table 7.17: Grid with graphics depicting mini-bus taxis.....	252
Table 7.18 Example 1.....	254
Table 7.19 Example 2	257
Table 7.20 Example 3	258
Table 7.21 Example 4	260

LIST OF FIGURES

Figure 4.1: Model of the assessment instrument development process	98
Figure 4.2: Components of language competence	108
Figure 4.3: Components of communicative language ability in communicative	110
Figure 4.4: Weir's socio-cognitive framework	115
Figure 4.5: Validation framework designed by Shaw and Weir.....	116
Figure 5.1: Example fit statistics report for new scale trial Model 2	169
Figure 6.1: Distribution of items and persons: existing scale, Markers 1 – 6 (60 scripts)	185
Figure 6.2: Distribution of items and persons: existing scale, Markers 2 - 6 (60 scripts)	187
Figure 6.3: Distribution of items and persons: existing scale, Markers 1 – 8 (30 scripts)	188
Figure 6.4: Distribution of items and persons: existing scale, Markers 2 – 8 (30 scripts)	189
Figure 6.5: Distribution of items and persons: existing scale, Markers 1, 6, 9, 10 (30 scripts).....	190
Figure 7.1: Model 1, 60 scripts, 5 markers	232
Figure 7.2: Model 2: 5 markers, 60 scripts	243

CHAPTER 1: INTRODUCTION

1.1 RESEARCH OVERVIEW

The title of the thesis is: *Validation of a rating scale for distance education university student essays in a literature-based module*. In this chapter, the topic has been placed in context by presenting an overview of the research, including the background and rationale, followed by the thesis statement. This leads to a description of the research problem and the questions that were developed from the topic. This is followed by the aims and objectives of the study that were formulated to address the research problem. The research methodology chosen for this thesis is then discussed, followed by a brief account of the ethical considerations. Finally, the chapter provides an overview of the thesis in the form of an outline of the chapters. The points raised in this chapter are briefly summarised in the conclusion, in which the gap that was addressed by this research is reiterated.

1.2 BACKGROUND AND RATIONALE

Weigle (2002: 1) points out that the “ability to write effectively is becoming increasingly important in our global community, and instruction in writing is assuming an increasing role in both second- and foreign-language education”. The importance of effective academic writing is thus central to progress in tertiary education, where students are expected to master writing skills requiring an advanced standard of logical organisation and language accuracy, as well as reading skills which require an in-depth understanding of content. Valid assessment processes are of paramount importance in this context because, in order to maintain credibility and acceptable reading and writing standards, it is essential that rating scales should give an accurate, fair and balanced evaluation of the student’s ability. Therefore, assessment instruments should be valid, and validation should follow a rigorous process to ensure this.

Research in South Africa has demonstrated that academic writing presents a challenge to first-year students, particularly those studying in a distance education context in the

complex South African environment (Pienaar 2005; Spencer 1997, 1998, 2005; Spencer *et al.* 2005; Lephalala & Pienaar 2008; Chokwe 2011; Ward-Cox 2012). Furthermore, it would seem that students of content-based modules (such as the target module of this thesis, namely *Foundations in English Literary Studies* ENG1501) are inadequately prepared to deal with the level of critical thought and organisational skills required by these courses (Dovey 1994: 113; Butler 2006: 93). This lack of preparation is exacerbated by the distance learning environment, which is characterised by a lack of regular contact between students and lecturing staff, as well as minimal interaction between markers. These challenging conditions make this thesis unique and increase the importance of valid rating procedures for the purposes of both summative and formative assessment.

This research study was prompted by the researcher's employment as a tutor for ENG1501 at the Parow Regional Centre of the University of South Africa (hereafter referred to as Unisa), as well as the researcher's experience as a marker and e-tutor for other modules offered by the Unisa English Department. The ENG1501 course was designed for students to learn the basic principles of understanding and appreciating literary texts and the ability to write well-organised arguments in which they analyse and discuss issues raised by the texts (Appendix C). Based on the researcher's experience as a tutor, and evidence obtained from research conducted for her Master's dissertation (Ward-Cox 2012), it became clear that the writing ability of the targeted South African students is an area of great concern, particularly as many of them do not use English as their home language (Pienaar 2005; Spencer 1997, 1998, 2005; Spencer *et al.* 2005; Lephalala & Pienaar 2008; Chokwe 2011; Ward-Cox 2012). In South Africa, this situation is exacerbated by a school environment characterised by inadequately trained educators, poor infrastructure and constantly changing policies and curricula as evinced since the introduction of the now abandoned Outcomes Based Education (OBE) in 1995. Large and medium-sized studies of reading skills corroborate this evidence, and confirm the generally low literacy level of South African learners and students, especially those from disadvantaged socio-economic environments (Pretorius & Ribbens 2005; Mullis *et al.* 2007; Howie *et al.* 2008, 2016; Pretorius 2008; Pretorius & Currin 2010).

Against this background, not only is the teaching of writing and critical skills to the target group challenging, but the assessment of student writing assumes even greater importance in this specific context. This is because assessment in distance learning not only evaluates the student's ability for the promotion purposes (summative assessment), but also constitutes a main form of assignment feedback (formative assessment) given the minimal or, in many cases non-existent, face-to-face contact between students and markers. This is despite recent efforts to introduce online tutors, known as e-tutors (who generally are not the markers of the students' assignments). This inadequate formative function underlines the significance of accuracy and objectivity in the assessment process. Thus, it is essential to use a rating scale that clearly measures what it is supposed to measure, and where criteria are clear and unambiguous to all stakeholders. It was in this context that the concept of validity was investigated and the validation of the existing rating scale took place.

Furthermore, it would appear that, although there has been extensive research on feedback in Open Distance Learning (ODL), very little has been carried out in the area of empirical validation of rating scales in this context, particularly for first-year ODL modules that focus on literary studies (such as ENG1501). Thus, there was a need for research dealing specifically with the validation of a rating scale in the discipline of the target module to determine whether it met the appropriate criteria for summative as well as formative assessment in this module (and, by extension, in similar modules).

The rating scale used to assess the target group of this study was a modified version of that designed by Jacobs *et al.* (1981). Although the original scale had international credibility (Spencer 1998), the modified version (provided in Section 5.8) had not yet been tested empirically in the context of the target group. This might have meant that use of the scale was prone to marker subjectivity and interpretation, aggravated by the ODL context with its limited contact between markers.

With this problem in mind, this study was undertaken to investigate the validity of the existing scale used to assess academic writing in response to literary texts in an ODL environment. The results of the first validation process were presented to a panel of experts who decided on the steps to be taken to modify the existing scale in order to

produce an empirically validated scale for assessing the assignments of similar target groups.

It must be kept in mind that no rating scale, particularly in the Humanities, can be declared to be entirely reliable, and that no claim of this nature has been made for the rating scale developed or modified during the course of this study. However, it is believed that the new or revised scale will contribute to the ongoing quest for accuracy and objectivity in assessment.

1.3 THESIS STATEMENT

An empirically validated rating scale was developed to improve the accuracy, fairness and reliability of results for the formative and summative assessment of essays in the ENG1501 (*Foundations in English Literary Studies*) module at the University of South Africa (Unisa).

1.4 STATEMENT OF THE RESEARCH PROBLEM

Research has demonstrated that academic writing presents a challenge to first-year university students, particularly those studying in a distance education context (Pienaar 2005; Spencer 1997, 1998, 2005; Spencer *et al.* 2005; Lephalala & Pienaar 2008; Chokwe 2011; Ward-Cox 2012; Shandu 2017). Furthermore, it has been found that students of the literature-based modules are inadequately prepared to deal with the level of critical thought and organisation required by these courses (Butler 2006: 113). Furthermore, in distance education, the relative lack of regular contact between students and lecturing staff, as well as the paucity of interaction between markers, increases the importance of clear, valid assessment procedures. The problem is that there is a lack of research on how to ensure that a rating scale in the discipline of the target module meets the necessary criteria for summative as well as formative assessment in this and, by extension, in similar modules.

1.5 RESEARCH QUESTIONS

The primary research questions addressed in this study were:

1. Is the existing assessment scale used for the *Foundations in English Literary Studies* (ENG1501) at Unisa valid in terms of the various aspects of validation and purposes (namely formative assessment, summative assessment, feedback)¹
2. How can the existing scale be modified or replaced in order to produce a validated scale in terms of validity, including user-friendliness and inter-marker validity?

These primary research questions were supported by five sub- questions as follows:

1. What do the results of the empirical research process reveal about the validity of the existing scale?
2. What are the observations of the tutors and markers who use the scale to assess examinations and assessments for this module²
3. What effect, if any, does the distance learning, multilingual and multicultural context have on the perceived and actual validity of the scale?
4. What recommendations, principles and insights from other stakeholders can be employed to create an improved scale?
5. How can the modified or new rating scale be designed and tested to ensure optimum validity?

In summary, this study was undertaken to investigate the validity of an existing rating scale used to assess academic writing in response to literary texts in a distance education environment and to modify the existing scale with the aim of producing an empirically validated scale for assessing the essays of the target group. The problem thus addressed was the validation of a rating scale that was appropriate to its purpose and context.

¹ Note that, in this study, validity will be linked to reliability in line with recent research. Reliability is tested as an aspect of validity as discussed in Chapter 3.

² It was suggested that the thesis statement could thus read “ it is possible to create a rating scale for an ODL context which is easy to use, provides consistent inter-marker reliability valid for summative assessment and useful for formative assessment”.

1.6 AIM AND OBJECTIVES OF THE STUDY

In the light of the foregoing discussion, the aim of this study was to develop an empirically validated rating scale for assessing *Foundations in English Literary Studies* (ENG1501) assignments at the University of South Africa.

This aim was achieved by focusing on two areas as follows:

- evaluating the existing scale empirically;
- using the findings of the empirical process to modify the existing scale or develop a new scale.

In order to operationalise the aim of the research, a number of objectives were identified. These were to:

- examine the concepts of validity and validation;
- identify and describe an appropriate framework for the validation of the assessment of the assignments of the target module;
- evaluate the existing scale by examining examples of student writing, supplemented by questionnaires and comments from stakeholders;
- modify the existing scale or draw up a new scale (depending on the results of the process);
- validate the new or modified scale by means of quantitative and qualitative procedures;
- propose an empirically validated rating scale for assessing student assignments in the module *Foundations in English Literary Studies* (ENG1501) at the University of South Africa.

1.7 NATURE OF THE RESEARCH

A combination of quantitative and qualitative procedures was used to validate the rating scale. This is in accordance with the belief expressed by Bachman (2004: 6) that both

qualitative and quantitative approaches should be employed to establish the suitability of an assessment instrument in a particular context.

1.7.1 Theoretical underpinning

The initial chapters of the thesis contain a literature survey of the distance education context in which assessment takes place. These findings provided the background and context of the assessment and validation process. Research on assessment and the impact of multilingual, cultural and socio-economic factors on the assessment process in the ODL and open distance e-learning (ODeL) context and whether assessment in ODL changes the form and/or content of criterion-based assessment significantly are discussed.

The concepts of validity and reliability are examined in the following chapters. The discussion includes the differing opinions of the relationship between the two concepts, the relative importance of the two concepts as well as the various types of validity. In the subsequent discussion on validation, theoretical models and frameworks are examined in order to establish a foundation for the development and validation of an assessment scale that can be used to evaluate student writing. Furthermore, factors that influence scores, such as assessor training and characteristics, as well as inter-assessor differences, are taken into account. The theory provides the basis for the analysis and evaluation of the existing scale.

1.7.2 Research methodology

Hattingh (2009: 8) points out that empirical scale development “entails developing scales based on analyses of actual samples of learner writing. Such analyses may reveal typical traits of how the construct is manifested in practice.” The rating scale should incorporate a description of these traits. Hattingh (2009) adds that a further consideration of an empirical approach is the investigation of how criteria and descriptors will possibly be applied and interpreted by assessors. Thus, the process of validation includes various forms of evidence and combines quantitative and qualitative

methods in order to test and justify claims of validity (Weir, 2005: 15). In this study, the process was carried out according to the following steps:

Step 1: Testing the existing scale: collection and marking of scripts

For this research project, a random sample of 200 assignments, written by students registered for ENG1501, was collected by the researcher with the assistance of the module co-ordinator. The necessary permission to use the essays for research purposes was obtained (Section 1.8 and Appendix A1).

The assignments had been submitted online, so there was no need to photocopy or print them for the purpose of this study. The only change made was the deletion of students' names and registration numbers, and the addition of a randomly allocated script number in the case of a smaller sample (60 - 68 scripts) selected from the initial 200. This initial selection was made according to the original mark allocated to the scripts and represented all levels of the marking grid (Appendix B). The original mark was deleted from the scripts, which were then sent to markers, who, after a briefing session, marked them again, using the existing rating scale. Ten expert and experienced markers participated in this process. The group of assessors included markers and tutors of ENG1501 as well as lecturers and examiners of similar modules at other tertiary and secondary institutions. During a workshop, the group was familiarised with the existing rating scale, and shared an understanding of the assessment context and of the construct to be assessed.

Owing to practical considerations, it was not possible for each marker to assess all the scripts, so each one was expected to assess at least 30 essays, and each essay was assessed by at least five of the ten markers. Marks were assigned for content and use of language, as indicated by the scale, and markers were required also to allocate a combined score.

Step 2: Analysis of data

After the marking process, the data were statistically analysed, using the FACETS version of the multi-faceted Rasch programme (Linacre, 2006b) and correlation coefficients were calculated to supplement and verify the results obtained from the Rasch process. The purpose of the analysis was to examine

- scoring consistency among the markers;
- the degree to which the sample of essays represented the full range of student competencies on the scale;
- The accuracy of the levels at which essays were benchmarked by the assessors
- assessor bias
- the degree to which the rating instrument represented the construct under assessment.

The Rasch analysis is a valuable, multi-faceted procedure because it provides:

...conclusive documentation of the many ways in which rater behavior can vary, as well as ... identify some of the kinds of measures (such as training and multiple rating) that can be taken to assist in managing this variation (Lumley & Brown 2005: 830).

The Rasch reliability index uses data on various facets of the marking process (such as learner ability, assessor characteristics and item difficulty) in order to indicate the relationship between these facets, predict the student's likely score, and investigate the differences between levels of scores assigned by different assessors. Data can be used to demonstrate:

- item difficulty and assessor bias towards any of the features or criteria of the rating scale;
- the degree to which the features measure the same construct;
- the degree to which the features indicated on the rating scale reflect the construct being assessed.

The quantitative data were supplemented by qualitative information provided by questionnaires completed by tutors and markers as well as comments, notes and reports from the assessors. This feedback was analysed and provided a rich source of data.

Step 3: Modification of existing scale or development of new scale

The information gathered from the analysis of the marking process was used to revise and refine the assessment scale. This was carried out by a panel of five experts and included experienced educators and examiners of English at Unisa and other institutions. Communication was initially face-to-face and, thereafter, by means of electronic media such as email and Skype.

Prior to the initial one-day workshop, participants were briefed about the aims of the project and provided with background reading as well as examples of the benchmarked essays (representing different levels of competence), which they were requested to analyse. This preparation assisted them in discussing the efficacy of the rating scale.

During the workshop, the results of the quantitative and qualitative data were discussed in order to determine the type of scale that would be most suitable to assess the construct in the given context. The panel analysed the benchmarked examples of student writing to identify the salient features and distinguish performances at different levels of proficiency. The existing scale was closely examined to determine whether it met the necessary criteria and how it could be improved, revised or replaced. This led to the construction of a first draft of the modified/ new scale. The draft scale was then refined and revised, initially at the workshop and then during the course of subsequent meetings and/or communication among the panel members.

Step 4: Piloting of the revised scale

The planning and design process was followed by a trial of the revised scale during which the essays were scored by the markers in order to ensure reliability, and to evaluate the strengths and weaknesses of the scale. The results of this exercise were quantitatively and qualitatively analysed in the same way as the previous process in

order to modify and refine the evaluation approach posited by this research. Furthermore, results were discussed at each stage of the process. This provided rich information to reinforce the quantitative aspects of the data.

Thus, the steps followed in the research were:

- assessment exercise of the existing scale undertaken by a group of markers;
- data analysis;
- revision of the scale by a panel of experts;
- piloting of the new or modified scale by markers;
- feedback in the form of comments elicited by written feedback from markers, questionnaires sent to Unisa tutors, e-tutors, markers, module co-ordinators and moderators and in the course of general discussions at each stage.

The information obtained from the literature reviewed, as well as that extracted from the empirical data, was used to validate an assessment scale that met the needs of the target group.

1.8 ETHICAL CONSIDERATIONS

The researcher requested the consent of the participants (students, lecturers, markers, tutors, moderators and co-ordinators) who took part in the study. This was done by sending them a consent form (Appendices A2 – A4). This included a full and clear explanation of what was expected of them so that they could make informed choices to participate voluntarily (Terre Blanche & Durrheim 1999: 66). The consent form sent to students (Appendices A2 – A4) assured them of the parameters of the confidentiality of any information they supplied. In this study, no names, addresses or student numbers were used. Each script was allocated a number randomly, such as “Script 1”.

Research should never injure people participating in the study (Babbie & Mouton 2004: 522). For the purpose of this study, voluntary participants were not exposed to any

danger to themselves, their home life, work, friendships, community or any other connections. The students wrote the assignments in the context of their Unisa studies, in their own homes or places of work and study (i.e. wherever they chose to complete their assignments). It is noted that these assignments were part of the study programme and thus no extra work was required of the participating students.

Markers and staff members were also not exposed to any danger as a result of participating in the project. Most communication was carried out either electronically, in the environment of private homes, or at the Unisa Parow Learning Centre.

1.9 OUTLINE OF CHAPTERS

Chapter 1: Introduction

This chapter contains a systematic discussion of the research problem and context, and incorporates an introduction to the thesis, followed by background information on the purpose of the study and the research objectives. This background leads to the problem statement and motivation, which is clarified further by means of the research questions and discussion of possible research limitations. An overview of the research methodology and design, as well as the processes followed in data collection and analysis, is then provided. The chapter concludes with a brief summary of the contents of the chapter.

Chapter 2: The distance education context

This chapter commences with a definition of distance education (DE) or Open Distance Learning (ODL), as it is now termed. This is followed by a discussion of the difficulties and complexities associated with ODL, with particular reference to the South African socio-cultural and multilingual context. A theoretical framework is then provided by referring to research on language assessment in ODL in the South African and international context. Finally, the objectives and outcomes of the literature-based target module (ENG1501) and the possible implications of these with respect to the validation of a rating scale are discussed.

Chapter 3: Theories of validity

In this chapter, the theoretical research framework as well as a literature review of texts is provided. This includes the broad context and theory base, definitions of validity and descriptions of the types of validity such as criterion, content and construct validity. A discussion of the relationship between the various types of validity leads to the debate between traditional versus modern concepts of validity. The relationship between validity and reliability is discussed and the viewpoint of the researcher, based on arguments presented in the literature, is presented on these issues.

Chapter 4: The validation process

This chapter contains a description of a validation process and a discussion of various frameworks such as Weir's Socio-Cognitive Framework, the Cambridge ESOL Framework and Shaw and Weir's Interactionist Framework. The chapter continues with an overview of the design of a framework for validating a written assignment or essay. This includes a consideration of cognitive and context validity; characteristics of test-takers; relationship between writer and assessor; resources; knowledge of content; administration context and issues; physical conditions and constraints; security; scoring, cross-testing, consequential and scoring validity; washback effect and the avoidance of test bias. The chapter concludes with a brief overview of the points made.

Chapter 5: Research process and methodology

The research process and methodology are described in this chapter. This discussion includes a description of the benchmarking exercise, the assessment and grading of scripts and the quantitative and qualitative analysis of the assessment. Quantitative methodology such as statistical analysis (Rasch scales and the calculation of coefficients) and qualitative research such as written comments and questionnaires are described in detail. The chapter concludes with a description of the ethical considerations that were taken into account. A brief overview of the salient points of the chapter is given.

Chapter 6: Existing scale: Research procedure and findings

In this chapter, the empirical research carried out to determine the validity of the current rating scale is discussed, commencing with the selection and downloading of the scripts used in the research study, followed by an account of the pilot study employed to test

the process, after which the main study is discussed in detail. The quantitative and qualitative methods used to evaluate the results are then described. These include statistical analyses, comments from the markers involved in the research as well as questionnaires completed by Unisa e-tutors and markers of the module ENG501. The results are summarised and the main issues arising from the process are discussed. This leads to the next phase, namely the development and trial of a new rating scale.

Chapter 7: Development of the new scale

The process of developing a new scale is described in Chapter 7, based on the quantitative and qualitative findings discussed in Chapter 6. The composition of the panel and the subsequent panel discussions are given, and the process of developing a new scale (namely, the design stage, the construction stage and the trial stage) are discussed in detail. Evidence includes quantitative elements in the form of statistical analysis, as well as the qualitative features extracted from the comments of markers employed at various stages of the process. Finally, the new scale is presented and reasons given for this choice. The chapter concludes with a summary of the process followed and the results of this process.

Chapter 8 Recommendations and conclusion

In the final chapter of the thesis, the findings are consolidated and their implications are discussed. Answers to the research questions are given and explained with reference to the research process and findings. The limitations of the study are noted and recommendations for further research are made.

1.10 CONCLUSION

This chapter comprised an overview of the thesis, including the rationale, scope and value of the research, particularly in the context of tertiary distance education in South Africa. The thesis statement was provided and was followed by a discussion of the research problem, the research questions and the aims and objectives of the study. These elements formed the foundations of the research. The research methodology and ethical considerations were described within this theoretical framework. This was followed by an outline of the chapters of the thesis.

In conclusion, there is an evident lack of research dealing with assessment processes in literature modules such as the target module (ENG1501: Foundations in English Literary Studies), compared with language-based access courses in which writing skills only are assessed. While language competency is an outcome for all these modules, there are differences in emphasis, and it can be argued that the assessment instrument should be altered or adapted to reflect these differences. The validation of a rating scale to ensure that it measures what it ought to measure will contribute to the effectiveness and credibility of the assessment process in this module, and insights gained can be extrapolated to similar courses.

Striving for objectivity in any field of endeavour is an ongoing quest, and it would be unrealistic and pretentious to claim that total objectivity has been achieved in any activity, including assessment design. The researcher thus makes no claim that a new or revised rating scale is water-tight and resistant to subjectivity. However, it is hoped that this study will result in a valid, fair and reliable writing assessment scale that will measure what it is intended to measure according to criteria expressed as clearly and as unambiguously as possible, given the challenging multilingual and multicultural ODL context. It is envisaged that this study will make a useful contribution to research in the field of assessment, particularly in the discipline of literary studies in the distance learning environment.

CHAPTER 2: CONTEXT OF THE RESEARCH – TARGET GROUP AND MODULE

2.1 INTRODUCTION

In discussing the importance of context in the language testing process, Weir (2005: 18-19) states:

...language processing does not take place in a vacuum, so testers also need to specify the context in which this processing takes place. They need to provide... descriptions of the conditions under which these language operations are usually performed.

The target group of this thesis comprised first-year students registered for an English Literature module at Unisa. Thus, it was necessary to take cognisance of the distance learning context as the defining background of the present research.

This chapter proceeds from a discussion of the general background to the more specific focus of the thesis, namely, the validation of the rating scale with a view to improving and, if necessary, replacing it with a scale that is valid and appropriate to the target module and group (ENG1501 at Unisa). The chapter commences with an overview of the distance learning environment in general, and of language learning in ODL in particular, and includes references to the relevance of the research in the South African situation (which will be elaborated later in the chapter). The impact of ODL and ODeL is then considered briefly. This is followed by a discussion of assessment and feedback in the distance learning context, particularly in the South African environment. Information on the ethos of Unisa and on the demographic composition of the target group is included. The aims and stated outcomes of the target module are then interrogated with reference to the skills and diversity of the target group. The chapter continues with a consideration of literary studies and assessment against the background of research into the role of literature in education and the relationship between literature and language teaching. The conclusion of the chapter contains a brief summary of the issues raised.

2.2 DISTANCE EDUCATION AND LANGUAGE LEARNING

Distance education can be defined as:

Planned learning that normally occurs in a different place from teaching and as a result requires special techniques of course design, special instructional techniques, special methods of communication by electronic and other technology, as well as special organizational and administrative arrangements (Moore & Kearsley 2011: 2).

In this context, Vorobel and Kim (2012: 548) point out that the term “distance education” includes many types of learning and teaching. According to Gunawardena and McIsaac (2004: 358), types of distance education include open learning,³ distributed learning,⁴ and networked learning.⁵ The distance learning environment is being referred to increasingly as ODL, and now ODeL, and these terms are being employed by Unisa.

The characteristics of distance learning and of the factors that distinguish it from face-to-face learning (FTFL) are the subject of a vast area of research. For the purposes of this study, the scope was restricted to those aspects that were relevant to the development of an appropriate assessment scale for the chosen context, for use in formative (assignment feedback) as well as summative (final examination) assessment.

It was observed that, while many significant differences between the two learning approaches remain, the borders between distance learning and FTFL seem to be blurred, and many historically FTFL tertiary institutions increasingly are offering ODL courses and online student support in South Africa. For instance, the website of North West University (NWU) advertises online courses, describing open distance learning as “a delivery mode and teaching and learning approach that focuses on increased access to

³ Open learning can be defined as a type of distance learning with “open entry – open exit courses”. This implies the flexibility of the students’ schedule in terms of beginning and end of courses depending on students’ readiness and preferences (Gunawardena & McIsaac, 2004: 358).

⁴ Distributed learning can be defined as a type of distance learning which adopts a student-centred approach which allows flexibility with regard to place and time of study. To facilitate this flexibility, the course components are distributed across various media (Rennie, 2007).

⁵ Networked learning is defined as learning in which computer and information technology “is used to promote connections: between one learner and other learners; between learners and tutors; between a learning community and its learning resources” (Jones *et al.*, 2000: 18). Networked learning can be described also as learning in connected space (Gunawardena & McIsaac, 2004) .

education and training where barriers caused by time, place and pace of learning are eliminated” (distance.nwu.ac.za). The ODL teaching medium makes learning accessible to a wider student population that previously was deprived of educational opportunities owing to geographical distance, financial constraints and/or other challenges.

It is also possible that growing interest in ODL has been sparked, *inter alia*, by the disruption of classes at FTFL universities throughout South Africa as a result of the #FeesmustFall protests of 2015 and 2016. During these periods, many campuses were forced to close, and academic staff resorted to online communication, among other strategies, in order to assist students with the academic programme. Whatever the reasons, it would appear that elements of ODL are being adopted and developed by mainstream FTFL universities, and that the gap between FTFL and ODL institutions is narrowing.

Despite these developments, the difficulties facing distance learning remain problematic. In the context of language education, Ward-Cox (2012: 27) points out that language learning, “especially in adults, is highly complex”. Ward-Cox (2012: 27) adds that the problems arising from this complexity are “exacerbated in DE”. According to Ward-Cox (2012: 18), the particular challenges of distance learning include:

- the geographical distance between students and lecturers and students and peers;
- minimal face-to-face contact;
- the logistical and administrative problems that are not found in the FTFL environment.

However, although the lack of face-to-face interaction in the distance learning environment is generally perceived as an obstacle to learning, a different perspective is presented by Saba (2000). In a discussion of the difference between distance learning and FTFL, Saba (2000) argues that the lack of face-to-face communication in distance learning is not necessarily a negative factor. In distance learning, “interaction transcends the idea of distance in its physical sense and embraces the discussion of teaching and learning in general” (Saba 2000: 4). Similarly, Spencer (2009: 104) argues that

“[i]ronically, the increased enrolments and staff to student ratios, so characteristic of higher educational institutions world-wide, have a more negative impact in residential institutions than they do in their ODL counterparts”. This is because increased student numbers at FTFL institutions result in a directly proportionate increase in costs (for example, in terms of infrastructure) which, in turn, can cause “radical surgery to the volume of assessment and, in particular, to the volume of feedback” (Gibbs 2006: 12). On the other hand, the cost of assessment is not affected to such an extent at distance learning institutions, despite the fact that, at a university like Unisa, “enrolments for single courses are in the thousands” (Spencer, 2009: 104). The institution “simply hires more tutors” (Gibbs 2006: 13), or appoints more external examiners to the examining panel, in order to cope with an increase in the number of assessments submitted.

The observations of Saba (2000), Gibbs (2006) and Spencer (2009) are valid, provided satisfactory interaction takes place in the form of written feedback or online intervention. Unfortunately, in the case of the target group of this study, interaction was restricted owing to the constraints of the semester course (described in Section 2.5), and sometimes by delayed feedback. Face-to-face contact was possible in regional learning centres, but attendance at tutorial classes was voluntary and limited to those students who could access these venues. As mentioned in Section 2.3, the online e-tutor programme is a promising recent development which requires further empirical evaluation.

In a discussion of learner-centred language learning as applied to distance learning, White (2005; 2006) notes a move away from a “linear model based on fixed content” towards one with “fluid course elements which are developed through the contributions and interactions among learners and teachers, and the written and spoken texts they produce” (White, 2006: 251). White (2006: 251) envisages the ideal situation as one in which “learners both construct and operate at the learner-context interface, according to their own needs, preferences and beliefs and also in response to the demands and requirements of the learning context” (White 2005: 67). White (2005: 67) is of the opinion that students should develop self-knowledge, knowledge of the learning process, and knowledge of their environment. They should also attempt to adapt these to the exigencies and opportunities offered by the available distance learning programme or course, and even provide input into the structure of courses (White, 2006). It is,

however, unclear how the ideal environment described by White (2005; 2006) can be implemented within the stringent, one-semester timeframe required by modules such as ENG1501 at Unisa, although advances in online communication might offer a partial solution.

Solé and Hopkins (2007) agree with White (2005; 2006) that the central issue is the learner dimension, which incorporates the characteristics, needs, experiences and conceptualisations of the student. In a study of two distance language programmes at two tertiary European institutions, Solé and Hopkins (2007: 351) discuss the challenges faced by distance learning as the result of a pedagogical move “away from the cognitive models to more socio-constructive approaches to learning” in which “language learners assume a central role in the language learning experience”. Furthermore, the researchers observe that the student-centred approach emphasises “collaboration and interaction among learners”. However, a serious challenge is that of fostering and developing relationships in an environment in which the various parties are geographically distant from one another and represent a diversity of needs, viewpoints, experiences and cultural backgrounds (Fung 2017). This is the case pertaining to ENG1501, as can be seen from the diversity evident in the demographic details given in Section 5.6.1 - 5.6.3.

White (2005; 2006) and Solé and Hopkins (2007) concur that distance language learning faces the major challenge of meeting the students’ need “to develop knowledge of themselves, their learning processes, and the possibilities offered by their environment, and [to] try to integrate those with the distance educational possibilities available to them” (Solé & Hopkins 2007: 353). Solé and Hopkins (2007) emphasise the importance of student autonomy and meta-cognitive skills – defined as “those that relate to the individual’s previous experiences, self-knowledge and expectations for a particular learning task” (Solé & Hopkins 2007: 353) – in the implementation of the strategies and techniques necessary for successful learning.

Although the ideas of White (2005; 2006) and Solé and Hopkins (2007) are helpful in addressing the problems of language learning in a distance learning context, in order for the students to develop the necessary language skills, they need sufficient time to develop these skills. This implies that students are afforded time and opportunity to engage with the learning materials and form a relationship with the educators, despite

the disadvantages of the physical distance between stakeholders. Unfortunately, given the serious constraints of the ENG1501 module – a semester course which allows for only two written assignments and, therefore, little time for any intervention or exposure to the language of use and the requisite skills – it seems almost impossible for such ideas to be implemented without radically revising the timeframe of the course.⁶ In the present challenging situation, using a valid, clearly understandable, rating scale, at the least, will go some way towards the provision of meaningful feedback in a manner that is both accurate and practicable in terms of the constrained time scale.

Many of the factors affecting distance learning are shared by other fields of study in the distance learning environment, and research findings in these fields can be extrapolated to language learning in this shared context. For instance, the aim of a study by Wang *et al.* (2008) was to determine the interaction between learning, motivation, learning strategy and self-efficacy, and how these affected learning results. In the study, self-assessment questionnaires were distributed to 135 students (68 females and 67 males) enrolled at the Beijing Radio and Television University. These students were studying in a distance learning context and majoring in electronic information technology. According to the researchers, the findings demonstrated relationships between self-efficacy, learning strategies, and learner results. It is significant that positive learner motivation and effective learning strategies were found to correlate with positive and predictable results. In a South African context, the importance of assisting students to develop a sense of ownership of their work is advocated by researchers such as Spencer (1997; 1998; 2005; 2009), Spencer *et al.* (2005), Pienaar (2005), Lephalala and Pienaar (2008), Letseka (2016), Pitsoe and Letseka (2016) and Shandu (2017).

In a study of the difficulties associated with the tailoring of distance learning courses and feedback strategies to suit the individual needs of students, Thang (2005) surveyed Malaysian DE students' perceptions of English proficiency courses, particularly in respect of their opinions on the support and guidance received. The researcher obtained the information by means of a questionnaire and semi-structured interviews. Interestingly, it was found that students who participated in the interview claimed to want more support and guidance while those who completed the questionnaires desired

⁶ Fortunately, the duration of the ENG1501 course will be a year from 2020.

greater freedom. It was difficult to account for these differences, but it is significant that they indicate a range of student expectations and thus present a challenge in catering for individual needs. This challenge is exacerbated by the distance learning environment with its lack of personal contact and, by extension, the absence of face-to-face dialogue that could facilitate solutions to students' problems. Another challenge is to develop a strategy that gives sufficient learner support while encouraging the autonomy that many students appear to want, and which is an important component of successful distance study. Learning materials and intervention strategies will have to be developed to meet the challenge of balancing support from lecturers with the students' sense of ownership of their written work. In this regard, without claiming to provide a complete solution to a complex problem, using a clear and appropriate rating scale could assist formative assessment by giving guidelines and supplementing written feedback while at the same time fostering student autonomy by encouraging students to interrogate their own work.

2.3 OPEN DISTANCE LEARNING (ODL) AND OPEN DISTANCE E-LEARNING (ODEL)

New technologies have the potential to reduce the gap between distance learning and FTFL, and have been introduced into the Unisa learning environment in the form of online marking and online tutoring (e-tutoring). Thus, the distance learning context is increasingly being referred to as ODL and, latterly, as ODeL.

In an overview of current research articles on second- and foreign-language teaching in DE, Vorobel and Kim (2012) draw attention to the fact that an increasing number of international institutions offer courses which are either web-enhanced or completely web-based. This was emphasised by the research of Kramer (2008). The chief advantage of these courses, compared with FTFL, is their flexibility in terms of place, time and pace of learning. This flexibility allows educational access to students who cannot attend face-to-face classes for various reasons, including distance, venue and time (Gutske 2010).

Unfortunately, despite these obvious advantages, it appears that limited participation frequently presents a problem in e-learning. Cormier and Siemens (2010: 35) observe

that “learners now have considerably more access to content and more opportunities to engage online.... Yet analysis of the open courses ... reveals reluctance on the part of many learners to engage in open online discourse”. Cormier and Siemens (2010: 36) suggest that this is because of “strong personal reasons for not wanting to form an online identity through transparent open learning”, and cite privacy issues expressed by students.

It would seem, however, that the problem of this perceived lack of participation is more nuanced than that implied by Cormier and Siemens (2010: 35). In a survey of practitioners’ reactions to online tuition, Jones *et al.* (2000: 22) found that “low participation was the reported factor related to disappointment [in the courses]” but that “disappointment was a common but not a universal feature”. Not surprisingly, Jones *et al.* (2000: 25) noted that “practitioners who did not experience low participation did not express disappointment with course outcomes”. Jones *et al.* (2000: 25) found that, although the practitioners “expressed a similar common philosophy or paradigm, they did not have a stable repertoire of 'rules of thumb', of reliable design guidelines”. These would have been easier to develop in a traditional, face-to-face setting by means of, *inter alia*, participation in meetings, lectures and seminars, which, Jones *et al.* (2000: 25) state, “whilst not unproblematic, has a set of commonly understood assumptions”. In contrast, the “boundaries within a networked setting appeared less commonly understood” in online learning (Jones *et al.* 2000: 25). Jones *et al.* (2000: 26) add that these findings “raise questions of staff development and suggest that it required significantly more than simple training in the technology”.

In the overview of Vorobel and Kim (2012) the issue of student satisfaction and retention is examined and studies that compare distance learning in “various formats” (Vorobel and Kim 2012: 556) with courses offering FTFL, such as Harker and Koutsantoni (2005), Murday *et al.* (2008) and Young (2008), are investigated. For example, Murday *et al.* (2008) found that online foreign language courses yield greater student satisfaction over time than the FTFL equivalents. However, in a comparison of online and blended formats in English for Academic Purposes (EAP) course, Harker and Koutsantoni (2005) conclude that the blended format was more effective than the purely online alternative with regard to student retention, although the results and satisfaction level were similar for both formats. Another study cited by Vorobel and

Kim (2012: 556) was that of Young (2008) who found that the effectiveness of language courses in DE formats greatly depends on the instructors' pedagogical effectiveness. This would appear to concur with the findings of Jones *et al.* (2000: 26), and points to the necessity for staff development that goes beyond simple "training in the technology".

Other researchers who have examined the effect (and effectiveness) of e-learning include Steeples *et al.* (2002), Hathaway (2015), and Cheng and Chau (2016) and Shandu (2017). Although the findings of these researchers differ somewhat with respect to the degree of effectiveness of the networked learning format, the researchers are in (qualified) agreement that this format has the potential to improve the quality of distance learning, provided it is implemented carefully.

Steeples *et al.* (2002: 323) caution against the seemingly indiscriminate use of the term "e-Learning" and warn that although:

e-Learning is a term that seems to have captured widespread support and enthusiasm, it is being used as a blanket term, in a variety of manners that are quite distinct from each other and, at worst, include a form of learning support is deeply concerning for the advancement of qualitatively rich and supportive learning experiences to people in higher education.

Steeples *et al.* (2002: 323) express concern about the proliferation of online courses offering what the researchers perceive to be "a quick way to get a degree or qualification", made all the more attractive to aspirant students because of economic factors. It would appear that the question of quality is secondary, at best, to cost and cost effectiveness in this context. Unfortunately, this is very likely to be to the long-term detriment of the students. Steeples *et al.* (2002: 323) point out that the "quick fix" approach that operates "at the level of transmission of information, providing little or no opportunities for the learners to engage with tutors or peers" will be to the advantage of neither the students nor the "long-term take-up of technologically-mediated forms of learning support."

While these issues were beyond the scope of this study, they point to the possibility of greater transparency in the assessment process as a result of enhanced pedagogical training and practice, a greater degree of communication with the students and, consequently, a better understanding of the assessment criteria, provided a solution is found to the problem of low participation by students in some instances. It can also be argued that a valid, accessible rating scale could help to alleviate the problems caused by the multi-faceted and very complex ODL context.

As Shandu (2017: 217) states:

It is not a matter of providing support but providing accessible and suitable support for those who need it most. What stands out from this journey is the power that an ODL institution has in changing the trajectories of people's lives ... where people could not have had an opportunity to study further due to their educational background as well as time and financial constraints, ODL provides opportunities.

However, Shandu (2017: 217) warns that these opportunities will be wasted if students do not receive the necessary support. She points out that “ODL principles should be embodied in cognitive, affective and systematic support intervention”.

2.4 ASSESSMENT AND FEEDBACK IN DISTANCE LEARNING

Describing the ethos of Unisa, Spencer (2009: 103–4) compares its founding principles with those of the Open University in the United Kingdom. Both institutions were founded on a “commitment to social equality” and an “ethos of inclusion” (Solé & Hopkins 2007: 354). Spencer (2009: 103) adds that “[i]n this context, a constructivist approach is unavoidable”. However, in practice, many modules offered at Unisa are dogged by administrative, logistical and time constraints which hinder a flexible and student-friendly situation.

Although instruction at Unisa is based mainly on “in-house-produced self-study language materials” which are designed to ensure that the “materials act as a surrogate teacher” (Solé & Hopkins 2007: 355), the diversity of the student body poses an almost

insurmountable challenge and, of necessity, the materials tend to represent a “one-size-fits-all” approach. More flexibility is possible in assignment feedback, which is provided by lecturers and, in the case of high registration numbers, by externally contracted markers. However, here too, the problem posed by the lack of face-to-face interaction arises and the fact that, in most cases, markers and students are not known to one another.

Feedback (including assessment results as an element of formative feedback) is particularly significant in the distance learning environment, since it often constitutes the only interaction between tutor and student. Hyland (2001: 233) points out that “interaction and feedback on performance are essential elements of the language learning process”. He adds that, since opportunities for face-to-face interaction in a distance learning context might be limited, feedback plays a central role in the dialogue between teachers and students. Hyland (2001: 233) examines the differences in the feedback of individual tutors and also the variations in “the type of feedback the students want and their reported uses of it”. These differences are exacerbated by distance learning where students have little or no direct contact with the tutor and find it difficult, and often impossible, to discuss needs, expectations, language difficulties and the interpretation of feedback.

A further problem is that feedback is sometimes delayed because of administrative and logistical problems. In the case of ENG1501, it is noted also that, in the majority of cases, the online e-tutor is not the marker of the student’s assignment, and that the same marker does not necessarily assess all the assignments submitted by a particular student. Thus, it is extremely important that assessment, including the rating scale, is reliable and that the criteria and feedback are unambiguous to all stakeholders in order to avoid inconsistency and contradictory interpretations that this could cause.

In addition, it is important to keep in mind that distance learning, with its lack of face-to-face tuition, poses a particular challenge to students, especially those whose home language is not English, which is the language of learning and teaching (LOLT) at Unisa (Section 5.6.1). This would seem to be one of the reasons why empirical evidence indicates that writing skills in particular pose serious problems for the target group, and that these problems are aggravated by the lack of regular interaction between

stakeholders (Spencer 1997; 1998; 2005; 2009; Spencer *et al.* 2005; Pienaar 2005; Pienaar & Lephalala 2008; Chokwe 2011; Ward-Cox 2012; Shandu 2017).

A further complication, as Du Plessis and Weideman (2014: 128) point out, is that “[i]n the South African school context, students are not necessarily first language speakers and the term Home Language (HL) (somewhat controversially) refers to the highest level of language instruction”. This implies that, in this context, the term “Home Language” is frequently a misnomer, as it might not be the language predominantly used by the student in personal and social contexts. Du Plessis and Weideman (2014: 128) add that “[i]t can thus not be assumed that spoken proficiency in a Home Language (HL) will be at the level of a first language, and even less so that writing ability will be on a high level.”

In addition to the issue of home language, the diversity of the South African student population gives rise to further challenges in teaching and learning (Section 5.6.1). This is true of the international context as well. In a research study at University College London (UCL), Fung (2017: 152) observes that in “an internationalised higher education sector, students bring very diverse prior experiences and expectations”. Basing his views on those of Levy and Petrulis (2012), Fung identifies five areas in which students might be challenged: “information literacy; personal beliefs about learning and knowledge; personal self-confidence; enquiry framing and direction setting; and peer collaboration” (Fung 2017: 152). These challenges are exacerbated by a distance learning context such as that of Unisa, and it can be argued that it is difficult (if not impossible) to achieve a satisfactory solution. However, attempts have been made to reach a partial resolution of the problems by exploring the communication possibilities offered by modern technology. In this environment, feedback that is consistent and comprehensible will contribute also to at least a partial solution to the challenges of distance learning, as listed by Fung (2017: 152), by providing a stepping stone to the eventual achievement of these implied goals.

The importance of feedback in formative assessment is further emphasised by Spencer (1997; 1998; 2005; 2009), Coetzee (2002: 139), Sieborger (2004: 11) and Lephalala and Pienaar (2008: 68). For instance, Coetzee (2002: 139) asserts that “formative assessment has a teaching, coaching and development function” and should thus “be

viewed as a process”. In this context, Sieborger (2004: 11) cautions that feedback should not be seen as “a final point of teaching and learning but [as] something which is used to guide and direct future teaching”. As has been mentioned, this would pose a challenge in distance learning with its lack of day-to-day interaction, although the study materials, online and written intervention (including a clear and accessible rating scale) should attempt to meet the criteria of guiding students and directing future teaching activities. As Lephalala and Pienaar (2008: 68) state, “[f]or feedback to be formative, its objectives should be aligned to the teaching and learning processes, and should meet students’ needs”. However, achieving this goal is a particular challenge in the very diverse South African distance learning context (Spencer 1997; 1998; 2005; 2009; Pienaar 2005; Spencer *et al.*, 2005; Lephalala & Makoe 2012; Ward-Cox 2012). An approximation of the ideal situation would require a careful alignment of assessment to the given construct, supplemented by increased and more effective communication strategies between lecturers and students. This could be achieved possibly by means of formal needs analyses and surveys, as well as more informal two-way communication between students and lecturers or tutors, using electronic media, *inter alia*. As Solé and Hopkins (2007: 353) point out, “assessment must be congruent with and closely reflect the course materials and skills taught during the course”. This alignment should also be made clear to the students, possibly in the learning material, further clarified by the online tutorials and by assignment feedback.

Research in a South African distance learning context thus emphasises the importance of appropriate assessment and feedback. For example, Spencer (1997: 48) recommends that feedback should contain “useable information on the strengths and weaknesses” of the text, and that marks awarded should “provide incentives and opportunities for improving performance”. Spencer (1997: 46 – 47) maintains that lecturers should change their approach “towards a form of assessment which is not restricted to monitoring, but aims to improve performance”. In an unpublished doctoral thesis, Spencer (1998: 10) believes that “response is only as effective as the student’s ability to grasp what has been conveyed, internalise the knowledge, and use it constructively in the learning process”.

In a further article, Spencer (2005) reinforces her previous research and describes taxonomy of tutor commentaries in response to student writing in a tertiary DE context.

One of the most significant findings was that there was a disproportionate emphasis on form as opposed to content found in tutor commentaries (Spencer, 2005: 220). In order to counter this tendency, Spencer (1998, 2005) makes use of the marking grid (Appendix B) adapted from Jacobs *et al.* (1981) in *ESL Composition Profile*. The important distinction which the marking grid makes between errors that do not affect meaning, and those that obscure meaning aims to prevent a focus on form to the detriment of meaning. However, it should be borne in mind that form should not be ignored, especially in the context of university education where students are required to express themselves clearly and accurately in the language of teaching and learning (LOLT). This raises the contentious issue of multilingualism and multiculturalism and the extent to which features of South African English (SAE) should be accepted. This delicate balance presented a challenge that the current study attempted to address.

Lephalala and Makoe's (2012: 2) study with Unisa students stresses the importance of recognising "the impact of culture and society on learning development", and add that it is therefore essential for distance learning institutions to have "an understanding of, as well as embrace, their students' socio-cultural contexts, in order to deliver educational programmes that are responsive to their students' needs". Lephalala and Makoe (2012: 2) conclude that "access to higher education can only be successful if distance education providers understand the varying contexts and socio-cultural circumstances of their students". While this noble sentiment is to be applauded, it raises questions pertaining to its practical applications. These include whether easier access implies the lowering or alteration of current entrance requirements and the extent to which Eurocentric references (as exemplified in the canon of literary works) should be included or restricted in favour of African-authored texts in English. Furthermore, it is difficult to understand how such an implied comprehensive understanding can be achieved in practice, given the diversity of the student population at Unisa (Section 5.6) and the time constraints on the markers and course designers (as discussed in Section 2.5). However, it can be argued that an awareness of cultural differences is essential and should be cultivated as far as possible.

In a foreword to the study of the philosophy of Ubuntu⁷ in the ODL context, Makhanya (2016: vii) avers that the “question of optimising equity in terms of access and outcomes is critical in South Africa”. According to Makhanya (2016: vii), this is because:

ODL has the capacity to bridge geographic divides and make connections. It brings people together in different communities of practice that underpin the increasingly important pedagogy of learning ecosystems, where students work together and learn together, sharing and caring for one another.

Makhanya (2016: vii) believes that, in the same way that Ubuntu philosophy stresses “the interconnectedness of people... whilst simultaneously recognising uniqueness and difference”, the ethos of the ODL environment “creates a space for a learning community to be built but allows each learner to retain his/her identity and learn at his/her own pace from wherever he/she sits.”

Although Makhanya (2016: vii) is referring to a specifically South African environment, his views are similar to those of Macdonald and Pinheiro (2012: 89-90). In a study of grammar teaching in schools based on the ideas of Vygotsky’s Socio-cultural approach, Macdonald and Pinheiro (2012: 89-90) state that to “isolate the child from her larger situation and to disregard the importance of the social in the learning process is to render learning a meaningless and fruitless task”.

It can be argued that the ideal situation sketched by this interactive, socio-constructive approach is challenged by the problem of implementation in the current South African distance learning environment. Mahlangu (2016: 111) believes that stringent quality assurance and, particularly, effective assessment practices will play a vital part in creating the desired ethos. Mahlangu (2016: 111) observes that: “Different outcomes and conditions must be well-defined operationally so that they can be measured. This entails the description of performance standards and criteria.... In any measurement exercise, the measure must be considered to be valid, to assess what it claims to measure, and to be consistent by producing reliable outcomes”. Mahlangu (2016: 111) claims that quality assurance based on the philosophy of Ubuntu “safeguards the ODL

⁷ According to Makhanya (2016: vii), “Ubuntu is part of the African philosophy which strongly emphasises the interconnectedness of people. It speaks to the ethic of humanity as part of a collective, whilst simultaneously recognising uniqueness and difference.”

institution's mission and aims that are clear and known to all. It ensures that everyone's duties are clearly stated and understood by all involved. It explains and documents the ODL institution's sense of 'quality', namely to check that everything is working according to plan".

These considerations were borne in mind during the examination of the existing assessment rating scale in order to ascertain whether it did in fact measure what it claims to measure in the particular context of this study. As discussed in Chapter 3, it is essential that the chosen scale provides fair, accurate and objective assessment of the written work of the target group, and it can be argued that fairness incorporates (as far as possible) an awareness of the socio-cultural environment and the potential consequences of assessment in that context.

In their re-interpretation of Messick's Validity Matrix, McNamara and Roever (2006: 14) emphasise the importance of context and its relation to fairness, which is achieved by "using evidence in support of test claims". This evidence takes into account the impact of "social and cultural values and assumptions" that "underlie test constructs and the sense we make of test scores" (McNamara & Roever 2006: 14). Messick (1989: 18) believes that:

For a fully unified view of validity, it must also be recognised that the appropriateness, meaningfulness, and usefulness of score-based inferences depend as well on the social consequences of testing. Therefore, social values and social consequences cannot be ignored in consideration of validity.

These issues have been explored in the discussion of validity theories (Chapter 3). What is noted here is the challenge of attempting to meet these criteria in the complex multicultural, multilingual Unisa environment that includes tertiary students from various language groups and a large variety of backgrounds, and the role of assessment and rating scales in this endeavour (Section 5.6).

Challenging though the ideal of achieving fair and accurate assessment might be, especially in the distance learning milieu, it is essential to strive towards this goal since the stakes are high for students attempting to obtain a tertiary education, particularly in

the South African environment. Fair and accurate assessment is important in respect of the summative function of the assessment, as the final result can have a far-reaching impact on the student's academic and, ultimately, socio-economic future. However, it is equally important to consider the formative impact of assessment feedback, especially its role in guiding the student and encouraging accuracy, fluency and the development of organisational skills. In order for these skills to be fostered, the student needs to be able to trust, understand and become familiar with the feedback and assessment criteria which, by implication, should be clear, balanced, consistently applied and unambiguous.

As has been noted, the relative lack of regular contact between students and lecturing staff, as well as the limited interaction between markers, renders direct interaction difficult, but it also increases the importance of clear and reliable assessment procedures, understandable to all stakeholders in a relatively communication-poor environment. At the very least, there should be written feedback regarding the rating scale from staff and students with a view to possible improvements. This was one of the aims of the current study.

2.5 THE RATING SCALE AS FEEDBACK IN FORMATIVE ASSESSMENT

McKenna (2007: 22) states that “[s]tudents, particularly those at first year level, are often unaware of what assessment practices are valued in higher education. Assessment rubrics are one means by which lecturers can make clear to their students what is expected of them before they undertake the task”. McKenna (2007: 22) adds that students “are frequently uncertain about what is expected of them in an assessment”, and consequently are often disappointed in their results. This gap is exacerbated by the distance learning context, which, as has been noted in Section 2.4, is characterised by limited contact between stakeholders.

McLoughlin (2001: 19) observes that there is often a discrepancy between students' and teachers' perceptions:

Students see the core activity of teaching as assessment. While teachers see it as teaching activity, culminating in assessment, students will define learning outcomes according to the types of assessment tasks they complete.

McLoughlin (2001: 19) avers that “a match between assessment tasks, learning activities and objectives will result in constructive alignment i.e. the students learning what is intended by the outcomes of the course”. This opinion is affirmed by Spencer (2009: 104).

With regard to feedback to students’ writing, Pienaar (2005) emphasises the need to “empower the students to take greater responsibility for their writing” (Pienaar, 2005: 193). She suggests that this can be achieved by self-correction and editing, which also leads to students familiarising themselves with the assessment scale. In the study by Pienaar (2005), students enrolled for Unisa’s English for Academic Purposes (ENN 103-F) course were given criteria to assess their own writing before editing it. The accuracy of their changes was then confirmed by the lecturer. Pienaar (2005) believes that assessment should not be merely a monitoring function but should aim “to improve the students’ performance” (Pienaar 2005: 194) by showing them “how to make connections between the feedback and the quality of their work and how to improve their writing for future assignments” (Pienaar 2005: 201). Pienaar (2005: 201) avers that, “when [students] become familiar with the assessment criteria, they will use the information to judge their own work”. Similar to Spencer (1998, 2005), Pienaar (2005: 202) makes use of the rating scale adapted from *ESL Composition Profile* by Jacobs *et al.* (1981).

Unfortunately, with the introduction of the semester system at Unisa, it is difficult to adopt these ideas owing to the constraints discussed in the previous sections. The following schedule for ENG1501 gives some indication of the very limited timeframe within which the module operates.

Registration takes place in late January to early February (first semester) and mid to late June (second semester). Registration might be disrupted by protest action, in which cases these deadlines may be extended. This has an impact on the due date for Assignment 1, which may also have to be changed to a later date.

In 2016, due dates for assignments for ENG1501 were given as follows:

Assignment 01:

2016 March 09 (first semester)

2016 August 31 (second semester)

Assignment 02:

2016 April 13 (first semester)

2016 September 28 (second semester).

The examination takes place in May/June for the first semester and in mid-October for the second semester, effectively creating a four-month ‘semester’ system.

In view of this schedule, it can be seen that there is a danger that a “crash course” ethos (which does not allow for editing and reflection) could evolve, to the detriment of the development of the students’ writing skills. It might be possible, however, to mitigate these difficulties to some extent by encouraging students to engage with an accessible rating scale.

Based on research findings, Spencer (2009: 109) points out that “learners in a distance-teaching context can improve both the content and the formal aspects of their writing without tutorial intervention; by simply being required to re-write and by being given a comprehensive guide to self-assess their work”. Spencer (2009: 109) adds that the benefits can be increased if “these ... strategies are combined by requiring both revision and guided self-assessment input” and stresses that the “benefits of self-regulated learning cannot be overemphasised in a distance-teaching context where lecturer feedback is challenging as a result of the high registration figures, the delay between submission of the assignment and tutorial feedback, and the difficulty of maintaining inter-rater reliability when an extended marking panel is employed”.

Spencer (2009: 105) suggests that training should take place “so that students can internalise standards and independently judge their work against the listed criteria”. Spencer (2009: 105) adds that this self-monitoring would benefit the instructor and the

student and lead to improved “personalised, individualised learning” that is recognised as “a challenge in an ODL context where scripts arrive in the thousands, identified only by a student name, number and address” (Spencer, 2009: 105). The situation as regards numbers remains challenging, with student numbers for ENG1501 cited as being 9383 and 7258 for the first and second semesters of 2016 respectively.

However, Spencer (2009: 109) reminds us that “assignment tasks form part of a much larger learning context” and that “no matter how carefully constructed any single assessment task is, it represents only one aspect in a broader educational context and it is the whole educational environment, rather than a single part, which needs to be optimised”. While it is acknowledged that there is no quick or single solution to this extremely complex problem, these conditions emphasise the importance of a valid, easily understandable rating scale which students can be encouraged to use for self-assessment, even in the absence of formal editing exercises. Although this is not intended to reform the overall environment, it does signal a positive development.

In the case of formative assessment, a clear and appropriate rating scale could supplement the learning materials and “act as a surrogate teacher” (Solé & Hopkins 2007: 355) by functioning as a scaffold to improve students’ understanding and foster an eventual sense of ownership of their work. However, the caution by Lamb and Simpson (2011: 51) not to “assume too much prior knowledge” on the student’s part should be heeded. Lamb and Simpson (2011: 51) argue that such a flawed expectation would be “reflected in the students’ frustration”. Lamb and Simpson (2011: 51) recommend strongly that “feedback should begin with students’ previous experience of assessment which invariably is the more scaffolded approach they receive at school”.

In a survey of students registered for courses in Geography and Engineering at the University of Johannesburg, Simpson and McKay (2013: 25 - 25) found that “the use of rubrics is a socio-cultural proficiency which is developed over time”. They add that:

The most significant finding to emerge from the results... is that, while the rubrics had gone some way towards making criteria for academic writing explicit, they are complex artefacts which require a great deal of brokering and, as such, may be opaque rather than transparent.

Simpson and McKay (2013: 25) noted that, despite the self-assessment exercises “students’ feelings about assessment rubrics and their use” remained ambivalent. This gave rise to the question whether “the idea that rubrics necessarily capacitate students and increase their confidence in approaching assessment tasks may need to be revisited”. These negative findings are shared by Trofimovich *et al.* (2014: 5), who found a weak correlation between self- and other-assessment although, in this case, in the context of English Second Language (ESL) oral assessment.

However, in the course of their research, Simpson and McKay (2013: 25) noted a positive finding that the alignment of marks improved with practice and that “the fact that [the students’] revised marks were more closely aligned with the score awarded by the lecturer suggests a developing understanding of the expectations contained in the assessment rubric”. This was illustrated by the fact that more (and more elementary) problems were experienced by the Geography students (who were first-year students) than by the Civil Engineering students (who were in their fourth year of study). Simpson and McKay (2013: 25) concluded that this contrast between the two groups suggests that the “basic conventions of the assessment rubric” explained or “unpacked” before students can engage meaningfully with it.

Furthermore, as Simpson and McKay (2013: 25) point out:

When students enter into the assessment experience, they do not enjoy the shared socio-cultural system necessary for effective learning to take place. Instead, this shared sociocultural system, or repertoire, **needs to be brokered by lecturers**, using the boundary artefacts they have created, such as the assessment rubric [my emphasis].

The readiness (or lack thereof) of students for tertiary education has been discussed in more detail in Section 2.7, but it is noted at this point that simply suggesting to students that they engage with the rating scale is insufficient for the scale to function as an effective formative instrument. Effective engagement should be viewed as a guided process, and not an instant panacea.

Solé and Hopkins (2007) also note that the problems of language learning are more acute in the acquisition of oral skills in distance learning than in the case of writing skills, which can be practised and assessed relatively easily by the use of post, e-mail or in a virtual writing environment. However, Solé and Hopkins (2007: 355) acknowledge that, in the distance learning environment, “providing opportunities for peer learning, one of the main tenets of socio-constructivism is... a challenge”. This challenge is also present in the assessment of written work owing to a lack of direct communication. Solé and Hopkins (2007) note that assessment should closely reflect the course material (content validity), and the tutors’ intentions should be clear to the students. This belief can be extrapolated to include the assessment scale that should provide a clear and accurate reflection of the construct being assessed. Solé and Hopkins (2007: 353) suggest the fostering of “meaningful dialogue between tutor and student”, as well as the effective use of “new technologies” (Solé & Hopkins 2007: 354). These can include communication by social media and, in particular cellphones, to interact with students. Specific suggestions for using these media for formative assessment are made in Chapters 6, 7 and 8.

2.6 Target module

According to the Module Form (Appendix C), the aims of the ENG1501 module are to “establish a literary and academic foundation for English Studies” as well as to “introduce students to representations of diversity in literature”. The outcomes are summarised as follows:

Students credited with this module will be able to apply appropriate reading strategies to a wide variety of literary and non-literary texts in English. They will also be able to demonstrate basic skills of writing academic English.

Two specific outcomes are extrapolated from this general statement. The first is to “read a range of literary texts in different genres (poetry, prose and drama) with comprehension at an inferential level”. The criteria for both formative and summative assessment for this specific outcome are as follows:

- A selection of literary texts is read and commented on, using acceptable academic discourse.

- Accepted conventions of academic discourse are applied.

The second specific outcome is to “demonstrate basic awareness of the creative choices made by writers of literary texts in English”. The assessment criteria apply to formative and summative assessment and are stated as follows:

- The dimensions of artistry and contrivance in the composition of literary texts in English are explored and explained through acceptable academic discourse.
- Accepted conventions of academic discourse are applied.

It can be argued that wording such as “the dimensions of artistry and contrivance... are explained” and “acceptable academic discourse” are vague and thus open to misinterpretation and subjectivity. These statements would need to be clarified, possibly by means of consensus among the teaching team and, subsequently, the markers, or at least by means of written communication to them. The assessment scale would have to reflect these more detailed, clarified, specific criteria. The dual purpose of assessment, namely its formative and summative functions, is nevertheless clearly stated in the outcomes, and this underlines the importance of valid scoring criteria that can be used for both of these functions.

The guidelines (2015) given to markers of the summative assessment (final examination) are stated more simply as follows:

The two crucial criteria for passing examination questions are as follows:

- a) the candidate’s ability to address or answer the question; and
- b) to do so in correct, standard English.

In summary, the two areas covered by these outcomes are the ability to:

- read a range of texts with insight and discernment, as well as to be able to identify and discuss stylistic and technical features of the text; and
- write about these texts using basic academic discourse.

It was decided to define the construct in these terms for the sake of the current research although, in practice, more clarity would be desirable in terms of the definitions given in the outcomes.

As shown in the Module Form (Appendix C), the target module (ENG1501) forms part of first-year English Studies, frequently studied in conjunction with a language-based module, *Introduction to Applied English Language Studies* (ENG1502). Although the two modules are presented and assessed separately, they can be regarded as being complementary, and indicate an integration of language and literature in certain aspects, such as the analysis of texts.

In reviewing the outcomes, there appears to be a gap between the skills level of students and the aims and expectations of the course. The issue is whether and to what extent this discrepancy should be addressed in the rating scale. In order to do this, reference is made to the current state of reading and writing skills of the target group.

2.7 READING AND WRITING SKILLS IN THE CURRENT SITUATION

Bearing in mind the assertion of Lephalala and Makoe (2012: 2) that it is essential for distance learning institutions to take cognisance of the socio-cultural contexts of the students, the current situation must be examined with a view to delivering “educational programmes that are responsive to... students’ needs”. Assessment practices form part of these programmes and, thus, a discussion of the current situation regarding reading and writing skills was undertaken in this study.

Reading and writing skills have been shown to be problematic for target groups similar to those who enrolled for ENG1501 (Spencer *et al.* 2005; Pretorius 2005; Butler 2006; Pienaar & Lephalala 2008; Chokwe 2011; Ward-Cox 2012), and it is questionable whether many of the students meet the pre-requisites as found in the outcomes statement as follows:

The following levels of learning ought to be in place to ensure successful completion of this unit standard:

The credit calculation is based on the assumption that students have successfully completed Grade 12 and are already competent in terms of the following:

- the ability to read texts in a focused and critical way;
- the ability to communicate information coherently and reliably in the language of tuition using basic conventions of academic discourse;
- the ability to take responsibility for their own learning in a distance learning environment.

In a study of first-year students studying an integrated literature and language course at North West University, Butler (2006: 93) remarks that the “extremely low competence of many students at entry level meant that meeting their real academic needs and maintaining standards commensurate with study at tertiary level were often incompatible aims”. This situation illustrates Elton’s (1993: 138) reference to the “double loyalties” of lecturers: “to their discipline, which represents what they teach, and to university pedagogy, which represents whom they teach”.

2.7.1 Reading skills

Academic language skills are regarded as a primary indicator of academic success in tertiary studies and this success is largely dependent on students’ initial level of literacy (Weideman & Van Rensburg 2002; Ntuli & Pretorius 2005; Van der Slik & Weideman 2008; Wildsmith-Cromarty & Steinke 2014: 38-39). Reading plays a crucial role in this regard.

In the case of the target module, “the ability to read texts in a focused and critical way” is stated as a criterion in the outcomes of the module (Appendix C). Students are expected to be able to read the prescribed literary texts with insight and discernment, as well as to be able to identify and discuss stylistic and technical features of the text. Unfortunately, as Wildsmith-Cromarty and Steinke (2014: 38) note, “the South African Education system continues to fail its children”. Wildsmith-Cromarty and Steinke

(2014: 38) point out that “ despite seventeen years of democracy, huge inequalities still exist between disadvantaged and advantaged communities in terms of teacher-training, teaching methods and physical resources”, and cite the findings of Machet and Pretorius (2008) that only 7.2% of schools in South Africa have functional libraries. This is disturbing, since Ntuli and Pretorius (2005) show a strong correlation between early print exposure and academic success.

Wildsmith-Cromarty and Steinke (2014: 38) also deplore the poor showing of South African schools in the Progress in International Reading Literacy Studies (PIRLS) of 2006 and 2011. In 2006, for example, South Africa was placed last out of 40 countries, and the scores demonstrate very little change in the 2011 report (Howie *et al.* 2012). The 2016 report also placed South Africa last out of 50 countries, and found that 78% of Grade 4 South African learners were unable to read for meaning in their Home Language (Howie *et al.* 2017).

Wildsmith-Cromarty and Steinke (2014: 38) claim that the problem is exacerbated by the widespread absence of a reading culture in the more disadvantaged socio-economic communities (Ntuli & Pretorius 2005) and has been compounded further by the National Department of Education’s failed attempt to implement the system of Outcomes Based Education (Jansen 1998). The unpreparedness of students entering universities has been confirmed by various lecturers specialising in academic literacy skills at South African universities (Weideman 2003; Boughey 2007; 2013; Steinke 2012).

Boughey (2013: 28) states that:

The assumption has always been that schooling prepares students for higher education. Once the notion of multiple literacies is acknowledged, then it becomes possible to identify school-based literacies that are different to literacies in higher education.

Boughey (2013: 29) agrees with Geisler (1994) whose review of research demonstrates a difference between literacy in schools and literacies in universities. Geisler (1994)

concludes that the assumption that schools prepare students for tertiary education is fallacious.

In view of this apparent disjuncture in the education system, the criterion stated in Outcome 2 (Appendix C), which requires the student to explore and explain the “dimensions of artistry and contrivance in the composition of literary texts in English... through acceptable academic discourse”, would appear to be virtually unattainable in many cases. The situation is similar to that described more than three decades ago in Asfour’s (1983) paper on the cultural barriers in teaching literature to Arab students at the University of Jordan. Discussing his students’ inability to identify differences in style and register, Asfour (1983: 80) stated that:

Unfortunately, our students simply do not have the opportunity to develop this sense of style on their own. The painful fact is that, due to various factors, reading is not yet a characteristic national habit, and the educational system in the country is not conducive to much extra-curricular reading on both the secondary school and university level.

This challenging state of affairs in the ongoing South African situation needs to be borne in mind when designing foundation courses for learning English at tertiary level and how they are assessed.

2.7.2 Academic writing in theory and practice

In an attempt to isolate “the typicality of academic discourse”, Patterson and Weideman (2013: 132) argue that the “typical analytical watermark or fingerprint is what sets factual academic texts ... apart from other texts created and produced in different kinds of discourse”. Patterson and Weideman (2013: 146) add that “academic language is the vehicle for verbalising the logically qualified process, in articulating the analyses and thoughts we organise in order to interact analytically with others”.

Blue (2003: 2) describes academic literacy as involving the sophisticated language ability that encompasses an “understanding of and ability to use appropriate disciplinary discourse”, as well as “the degree of autonomy expressed by the ability to criticise and

evaluate their own views and the views of others”. In practice, this ideal is a far cry from the language usage demonstrated by entry-level and first-year Unisa students registered for language and literature modules, whose assignments are frequently exemplified by elementary errors and minimal (or sometimes non-existent) organisational writing skills (Chokwe 2011; Ward-Cox 2012).

Hathaway (2015: 506) argues that “the recognition that international students need induction into academic literacy practices needs to be extended to all students, regardless of their linguistic backgrounds”. Hathaway (2015: 507) believes in “recasting and reframing the issues faced by all students in dealing with new literacy practices” and an “examination of the student experience in terms of their struggles to construct meanings in new discursual communities with attendant issues of identity and power, a process that can be more complicated and problematic than previously assumed”. It can be argued that this struggle is a problem faced by all first-year students to a greater or lesser extent. This can be extrapolated to the South African situation in which there are not as many international students, but a large number who do not claim English as a home language (Section 5.6.1.1). Bearing this in mind, one can argue that the recommendations made by Hathaway (2015: 507) are applicable to the South African context and that, similar to her target group, attention should be paid to all first-year students learning to engage with and produce academic discourse. Hathaway (2015: 514) adds that:

In fact, students are often painfully aware that academic language is both distinct from ‘everyday English’ and more difficult, that it is impersonal, formal, densely packed and full of specialist terms. This can cause anguish equally to home students and international students.

Hathaway (2015: 507) recommends an Academic Literacies approach that “focuses on acquisition of academic literacy practices of home students... and with a different emphasis” which “relocates the ‘problem’ away from the students and their supposed inadequacies to the task with its complexities and opaqueness”. This is a view shared by Lea and Street (2000) and Turner (2000).

In the South African context, Boughey (2013: 31) investigated “the implications for research and theory related to academic literacy in South African higher education” and makes the following observation:

... it is clear that understandings of the socially embedded nature of literacy have profound consequences for the ways in which we understand what students can and cannot do when they first enter university. Given the very different contexts in which they have been socialized, the ways they use language and the language related practices of reading and writing they engage in must be understood as involving more than mastery of what might be termed the ‘technicalities’ of language use.

In a study of first-year students at Unisa, Chokwe (2011: 139) concluded that the findings “unequivocally indicate that first year university students who participated in this study struggle with reading and writing (academic literacies)”. The problem is exacerbated by the distance learning context in which the research was conducted. Chokwe (2011: 139) states that the students’ writing is “fraught with grammatical mistakes, particularly with regard to spelling, sentence construction, tenses and punctuation”. Furthermore, students’ essays lack organisation (coherence) and structure. The students also struggle to formulate and present an argument. Chokwe (2011: 139) adds that it seems that “students learn academic writing for the first time at university”. Chokwe (2011: 139) argues that “if students struggle with the basic elements of writing, they will find it difficult to acquire other forms of academic literacies” and that:

The current state of student writing which is marked by poor grammatical correctness robs them and the academic staff of the opportunity to deal more with content and to fully integrate students to more academic literacies instead of being distracted by grammatical structures.

The findings by Chokwe (2011) confirm those of other researchers (Spencer 1997; 1998; 2009; Spencer *et al.* 2005; Pienaar 2005; Pretorius 2005; Butler 2006; Pienaar & Lephala 2008; Ward-Cox 2012). However, Chokwe (2011: 139) found also that “both students and tutors have similar ideas about what academic writing should entail” and that both groups agree that “feedback is an important part of teaching in ESL modules”.

In addition, students were aware of “problems with gaps that exist between high school and university”. Some areas of concern mentioned by students are:

- English proficiency;
- difficulties in collecting information;
- qualities of good writing;
- the need for more writing activities and “intensive” language support.

Most of these concerns will have to be addressed by tuition rather than by a rating scale but, as has been mentioned, the scale can be used in formative assessment to highlight features of academic writing that are important in academic discourse. Furthermore, the scale should be designed to prevent an over-emphasis on form at the expense of organisation and content. According to Spencer (1998; 2009) and Pienaar (2005), this has been achieved to a large extent in the *ESL Composition Profile* by Jacobs *et al.* (Section 4.5.2), used for the module under consideration in a slightly modified form. But the aim of the present research was to investigate these beliefs empirically and, depending on the results of the investigation, to introduce changes and/or refinements to this and other aspects of the scale.

2.8 LITERARY STUDIES AND ASSESSMENT

In interrogating the validity of an assessment scale for target groups such as that of the present study, cognisance should be taken of the current status of literature studies in the international and national arena. The decline in the traditional, privileged status of literature in favour of an approach emphasising the instrumental benefits of language competence is notable in international research and practice. For example, Dovey (1994: 288) favours an approach that focusses on skills rather than content, and does not grant literature its previous status. Dovey (1994: 288) believes that the aim of English teaching is to enable students to become “competent speakers, readers and writers of English, and help them become critical interpreters of the various forms of language use they encounter in the world around them, and the range of texts which make up their culture”. Widdowson (1982: 7) describes the dilemma that this shift of emphasis has precipitated as “a question, posed from within, as to what English is, where it has got to,

whether it should have a future as a discrete discipline, and if it does, in what ways it might be reconstituted”.

The changing attitudes towards the study of literature has been traced by Durant (1993: 158 - 160) in a three-phase model. As Butler (2006: 46) points out, “Although the model was conceived from the point of view of the language teacher... it is equally valid from the perspective of the development of literary studies”. According to Durant (1993: 158 - 160), the first phase of the model reflects the traditional view that sees the study of literature as an end in itself. This point of view is no longer widely held and seems to be increasingly difficult to justify. In the second phase, this belief is replaced by the more utilitarian viewpoint, such as that of Dovey (1994: 288), mentioned above. In the third phase the emphasis is on the role of literary studies in language acquisition. Here, the focus is on the practical benefits that the study of literary text has for the student, especially in the ESL context. In other words, literature is being seen increasingly as a vehicle for language learning and teaching, as well as a means of developing a personal response to texts. As Maley (1989: 59) declares: “Literature is back – but wearing different clothes”. The renewed interest in literature, which commenced in the last decades of the previous century, has been heralded by other researchers such as Widdowson (1983: 34), Hill (1986: 7), McRae (1991a: 432), Carter and Long (1991: 1), Falvey and Kennedy (1997b: 1), Paran (1998: 6; 2006b: 1), Prodromou (2000: 3) and Badal (2016).

Butler (2006: 14) agrees with McRae’s (1991b: 120) belief that language learning and literary study are “interdependent and, in a specialist context, should be seen as complementary at all stages in the educational process”. Butler (2006: 41) amplifies this view by pointing out that “Learners cannot develop literary competence without an adequate competence in language”. This is reflected in the outcomes for module ENG1501 (Appendix C) and in the current rating scale (Appendix B), which makes equal provision for content and language. However, in the case of ENG1501, which, unlike other first-year language modules, requires knowledge of a literary text (i.e. specific content), this equal rating needs to be interrogated and the relationship between the content and language re-evaluated. The outcomes given in Appendix C unfortunately do not give clarity on this issue, but the question should be asked whether

content should supersede language in this instance and, if so, how this relationship should be reflected in the rating scale.

As Butler (2006: 5) points out, the problems regarding the international trend of integrating language and literature are exacerbated in South African universities that previously catered exclusively for black students, during the apartheid era⁸ and which are described by Ruth (2001: 1) as “marginalised institutions usually in marginalised areas in a marginalised country on a marginalised continent serving marginalised communities”. In the case of historically white, tertiary institutions, the challenge is to cope with the changing nature of the student body, particularly a “growing enrolment of second-language, mainly black, students, who represented a challenge to the implicit linguistic and cultural assumptions of the departments” (Butler 2006: 5). This is the result of the separatist policies of the previous regime, which provided a markedly inferior education to black students, as well as severely limiting interaction between the various ethnic groups, particularly between the black and white populations.

Shanahan (1997) describes the conflicting aims of a university ESL course, namely developing linguistic competence on one hand while, at the same time, exposing students to a culture by means of literature. In this context, based on a study of student papers, Butler (2006: 65) notes that “there emerges a curious love-hate relationship with English: a recognition of its instrumental usefulness coupled with a profound ambivalence about the cultural baggage that the language and (more especially) the literature bring with them”. It can be argued that students, who are additional language speakers of English, are not interested in the culture of the L1 speakers of English. Kachru’s (1990: 3) concept of the diffusion of English “in terms of three concentric circles: the Inner Circle (L1 varieties, e.g. the USA and the UK), the Outer Circle (ESL varieties), and the Expanding Circle (EFL varieties)” could explain why members of the last two circles do not share the cultural norms associated with Inner Circle speakers of English, and have an instrumental rather than an integrative motive for learning the

⁸A segregated education policy was one of the central features of the racially separatist apartheid system. Unlike their White counterparts, Black students were subjected to a curriculum known as Bantu Education which concentrated on skills training rather than on academic subjects.

English language. This could be why literature in English by writers from the Outer and Expanding Circles is increasingly favoured for didactic purposes.⁹

In the case of Unisa, the situation is complicated by a number of first-language students who might have a higher level of linguistic competence as well as a less ambivalent attitude toward (and arguably greater familiarity with) the culture reflected in the texts. The prescribed texts for ENG1501 comprise an American (United States) novella, a South African novel, a South African drama and an anthology of poetry that consists of a selection of poems written by poets from South Africa, the rest of Africa, Britain, the United States of America and other countries where English is spoken. It should also be noted that the target module was not designed to be an ESL course, and that the aims of linguistic competence and exposure to the culture embodied by the texts apply to all students registered for the module. However, as has been seen, currently there is a majority of ESL students registered for the module (Sections 5.6.1. to 5.6.3) and these factors could give rise to many of the challenges and difficulties encountered. Furthermore, when prescribing the texts for this course, an effort was made to take demographics into account without ignoring the wider international context.

Another consideration is the fact that “The first year of university study has long been identified as being of crucial importance for its potential to provide a bridge to tertiary studies, particularly for educationally disadvantaged students” Butler (2006: 12). This statement reiterates the views of Dovey (1994: 268) who notes that it is “in the first year that a solid foundation must be laid for further study within the discipline, and students must be given something of value, which can be applied both within and outside the academy”. Dovey (1994: 268) adds that it is in the first year of study that “students experience the greatest difficulties, and it is at this level that the most significant changes are being made”. Regarding English Studies, Dovey (1994: 268) also mentions that many students require only a single credit in English, and have no intention of continuing into the second year. The English module therefore must serve a dual purpose. This is explained by Butler (2006: 113), who notes that the English Literature course is frequently prescribed as an “ancillary programme” for various programmes of

⁹ Kachru (1985: 12-14) does not list South Africa and Jamaica as belonging to any of these circles because of their “sociolinguistic complexity ... in terms of their English-using populations and the functions of English” (1990: 3). However, it can be argued that an ESL population has features of the Outer Circle.

study. According to Butler (2006: 113), this implies that the English literature module must therefore serve what Butler (2006: 113) refers to as a “double purpose”, on the one hand giving “grounding” to those – the minority – who will major in English literature, while simultaneously providing a self-contained course” that is both “interesting and useful to the majority”.

Butler (2006: 113) argues that the teacher is also required to “reach a (sometimes uneasy) compromise between meeting students’ practical linguistic needs and the demands of academic respectability”. Furthermore, Butler (2006: 113) ascribes the high numbers of first-year students, and the relatively few senior students enrolled for English Literature modules, to “two contrary forces currently operating in South Africa”. Butler (2006: 113) notes that while, on the one hand, the increasing emphasis on science and technology has resulted in low student numbers in social sciences and the humanities, including language departments, on the other hand, the relatively high number of first-year students registered for English Language (including literature) courses can be explained by the prestige enjoyed by English and its “position as a national lingua franca” in South Africa. This gives English an important instrumental role in South African education. However, there is an opposing view that English enjoys hegemonic status which contradicts the policy of 11 (equal status) official languages by unfairly advantaging L1 English speakers. This view also needs to be taken into consideration.

The implications for a valid and appropriate rating scale are that it should take into account various complex and possibly problematic factors, such as the stated outcomes upon which the construct is based, as well as the issue of fairness in the context of the target group, considering the complexity of the distance learning background, the possible consequences of assessment, and the variety of socio-economic and socio-cultural issues that impinge on the situation.

2.9 CONCLUSION

Based on the issues discussed in this chapter, not only is teaching writing and critical skills in distance learning challenging, but also the assessment of student writing assumes great importance in the specific context. This is because assessment in distance education not only evaluates the student's ability for the purpose of promotion (summative assessment), but also constitutes the main form of assignment feedback (formative assessment) given the minimal or, in the majority of cases, non-existent, face-to-face contact between stakeholders. This is despite recent efforts to introduce e-tutors (who are generally not the markers of the students' assignments). The formative function underlines the significance of accuracy, objectivity and fairness in the assessment process, factors which are also of paramount importance in summative assessment because of the academic and socio-economic consequences this final assessment might have for the student. It is thus essential that a rating scale clearly measures what it is supposed to measure, and that the criteria are clear and unambiguous to all stakeholders. It is in this context that the validation of a possible rating scale was undertaken. In this thesis the extent to which the distance learning environment and the specific outcomes of the target module affect the validation process has been examined. This process is dealt with in more detail in Chapter 4.

In this chapter, the context of this study (namely distance learning), and the importance of this context in assessment has been discussed. Particular focus was placed on research about language learning in ODL and the challenges arising from assessment and feedback in the distance learning context, particularly in the South African environment. The aims and stated outcomes of the target module were then described, and the gap between these and the skills of the target group was discussed. This was followed by an account of the current status of the study of literature in tertiary education, and the relationship between literature and language teaching. These issues provide the background for the following chapters on validity and validation, in which the focus has been narrowed to an overview of the current rating scale, and its perceived strengths and weaknesses, with particular reference to the need for a marking grid that supplies the pre-requisite information while, at the same time, promotes a quick turn-around time, taking into account the severe time constraints governing the module.

CHAPTER 3: THEORIES OF VALIDITY AND RELIABILITY: CHANGING PERSPECTIVES

3.1 INTRODUCTION

In this chapter, the theoretical research framework and the changing perspectives in the complex, constantly evolving concept of validity have been examined. The discussion includes theories and definitions of validity, and descriptions of the types of validity such as criterion, content and construct validity. This leads to the debate between traditional versus modern concepts of validity, and an evaluation of an interpretation of validity by Messick (1989). The concept of consequential validity is discussed with particular reference to its role in the practice of fair assessment. The link between validity and reliability is then examined and the viewpoint of the researcher, based on arguments presented in the literature, is presented.

3.2 THEORIES OF VALIDITY: CHANGING PERSPECTIVES

While it would appear that there is agreement on the general concept of the term “validity”, an exact definition has proved to be elusive, and differing perspectives have given rise to considerable debate. In order to achieve clarity on the issue, it is necessary to examine these ongoing debates and shifting perspectives of the definition of validity, and the relationship between the different types of validity.

3.2.1 Definitions of validity

Kane and Bridgeman (2017: 489) point out that:

General conceptions of validity grew out of basic concerns about the accuracy of score meanings and the appropriateness of score uses... and they have necessarily evolved over time as test score uses have expanded, as proposed interpretations have been extended and refined, and as the methodology of testing has become more sophisticated.

Kane and Bridgeman (2017: 491) trace a “gradual progression from simpler and more intuitive models for validity to more complex and comprehensive models”. This process will be described in this chapter.

Although there seems to be consensus that the term “validity” encompasses the ability of an assessment and/or assessment instrument to measure what it is supposed to measure, an exact definition of validity has proved to be difficult to formulate. While the general concept has remained stable for many years (Chapelle 1999: 254; Kane & Bridgeman 2017: 491), debates are ongoing about whether the commonly accepted definition is too broad (Kane 2004: 136), and how the various types and components of validity relate to one another. These differing interpretations point to the difficulty of analysing the precise nature of validity. A measure of clarity can be achieved by tracing the change in perspective from the traditional concept to the more unified interpretation adopted later by Messick (1989: 1992), while also taking into consideration the views of recent researchers interrogating this unified concept, for example, Weideman (2006; 2009; 2012), Rambiritch (2013), Van der Walt (2012); Chapelle (2012); Du Plessis and Weideman (2014) and Kane and Bridgeman (2017).

One of the first contributions about validity was by Cureton (1951: 621), who stated that “the essential question of test validity is how well a test does the job it is employed to do”. According to Cureton (1951: 622), “To be valid – that is to serve its purpose adequately – a test must measure something with reasonably high reliability, and that something must be fairly closely related to the function it is used to measure”.

Messick (1980; 1989; 1992) adopted a different angle, introducing an integrating concept of validity. Messick (1989: 13) described validity as “an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of the interpretation of the inferences and actions based on test scores and other modes of assessment”. As discussed in Section 3.2.1, Messick’s theory signalled a major change in the theories governing the relationships between the various types of validity, although it can be argued that he agreed with the general concept of validity, as can be seen in his use of the phrase “adequacy and appropriateness of the interpretation” (Messick 1989: 13).

The definition by Messick (1989: 13) shifted the emphasis to the inferences to be drawn from test scores, rather than the scores themselves. Weideman (2012: 3) notes that Messick's (1980: 1023) definition is further refined and reinterpreted by Kane (1992: 527), who states that, "Validity is associated with the interpretation assigned to test scores rather than with the scores or the test". Weideman (2012: 3) believes that Kane's "subtle reinterpretation of definition... lies in the emphasis placed... on the judgement of the adequacy and appropriateness of the inferences drawn from test scores ... , as against validity not being associated with either scores or the test". According to Weideman (2012: 3) the "fine" point of difference between the definitions of Messick (1989: 13) and (Kane 1992: 527) is that "Kane's redefinition of validity... speaks in the first instance only of an 'interpretation' that is assigned, not about its adequacy and appropriateness".

Later definitions are similar to that originally provided by Cureton (1951: 621). For instance, Weideman (2006: 74) states that "validity normally refers... to the power of a test to assess what it is designed to do". Hattingh (2009: 15) gives a similar description of validity, defining it as being "concerned with the question of whether a measurement instrument, such as a test or scale, measures what it claims to measure", whereas Van der Walt and Steyn (2007: 139) signal a shift in perspective by describing validity as concerning "an inherent attribute or characteristic of a test, that a psychologically real construct or attribute exists in the minds of the test taker". The central role of validity in assessment measurement is emphasised by Mahlangu (2016: 111), who describes the measurement of performance as being "concerned with defining the extent to which desired outcomes and conditions are being realized". Mahlangu (2016: 111) adds that in "any measurement exercise, the measure must be considered to be valid, to assess what it claims to measure, and to be consistent by producing reliable outcomes over time".

Thus, while there is broad agreement regarding the central function of validity, shifts in perspective can be noted in various definitions. These shifts reflect the differing views of validity theory adopted by various researchers. As Kane and Bridgeman (2017: 491) observe: "the major developments in validity theory have involved changes in what the term means and how it is used. The definition of validity has been and continues to be a work in progress".

3.2.2 Types of validity

In the first codification of validity standards, undertaken by the American Psychological Association (APA) (1954), validity was considered to be an indication of the degree to which a test could be employed to form a certain type of judgement on the test taker's performance (APA, 1954: 13; Shepard, 1993: 408). The study identified three types of validity, namely: criterion, content and construct validity.

3.2.2.1 Criterion validity

Criterion validity refers to the correlation of the test or assessment instrument with an external independent criterion, such as a test or rating scale that has been designed for the same purpose and in the same context. Hughes (1989: 22) points out that, in the case of criterion validity, the emphasis might not be so much on whether the instrument measures the construct, described by Hattingh (2009: 22) as “the relevant psychological structure that underlies a performance”, as on the extent to which the instrument is correlated with the external variable. Du Plessis and Weideman (2014: 132) explain that criterion-related validity “is established by correlating a test score with another measure of the same ability obtained at a different time”.

Criterion validity has two aspects, namely concurrent and predictive validity (Hughes 1989; Weir 2005; Fulcher & Davidson 2007). Concurrent validity indicates situations in which the assessment and the criterion are completed at the same time. In other words, concurrent validity would involve comparing a new test with an external criterion to determine whether it could be used as a substitute for a similar, existing test. Predictive validity is concerned with how well an assessment can predict a future criterion such as academic performance, as is the case in the Placement Test in English for Educational Purposes, or PTEEP.

It is essential that predictive validity is a reflection of the abilities of the test-taker because of the crucial importance that placement tests play in his/her future. Weideman (2006: 77) emphasises “the critical importance of predictive validity... i.e. whether the test can make more or less accurate predictions about the future success (or potential failure) of the performance of candidates who take it”.

Du Plessis and Weideman (2014: 132) note that it is “possibly because of the importance of criterion (and especially predictive) validity that for much of the twentieth century it was considered the paramount type of validity”. For example, Kane (2004: 137) notes that criterion validity was considered to be the “golden standard” during this period. Du Plessis and Weideman (2014: 132) explain that validity was measured “according to the degree of positive correlation of the assessment scale with a dependent variable”. This opinion was held by Guildford (1946: 429), Cureton (1951: 623) and, much later, by Shepard (1993: 409).

However, this view of correlation came under scrutiny in the early years of the twenty-first century. Kane (2004: 137) criticises the idea that the only measurement of the validity of an assessment instrument is its positive correlation with a dependent external variable. Kane points out that this interpretation is too wide and leads to the assumption that a measurement instrument is valid in relation to anything with which it correlates. The danger of this wide interpretation is that it could lead to an inaccurate interpretation of results. As Hattingh (2009: 17) points out, even if there is empirical evidence of correlation between the test or instrument and the external criterion, there is a risk of a false correlation if the criterion and the instrument demonstrate the same bias. In contrast, Messick (1989: 64) advocates a construct-centred approach instead of one based on the statistical analysis of the relationship between test and criterion scores. In Messick’s opinion: “There is simply no good way to judge the appropriateness, relevance, and usefulness of predictive inferences in the absence of evidence as to what the predictor and criterion scores mean”. This, as Kane and Bridgeman (2017: 519) mention, implies that what is of interest is the “relationship between the characteristics of test-takers and their future performance”.

These issues are of particular concern in the case of predictive validity because of its consequences for the future of the test taker (Messick 1989; McNamara & Roever 2006; Weideman 2006; 2012; Du Plessis & Weideman 2014). The implication is that false correlation of criteria, and the resulting test bias, could lead to decisions that impact negatively on the academic progress of the student and even, ultimately, his or her future socio-economic well-being (owing to incomplete qualifications and, therefore, fewer employment prospects). In this regard, a danger exists that tests could play a

gatekeeping function, serving to exclude students at an early stage in their academic careers. This will be discussed in more detail in Sub-sections 3.3.1 and 3.3.2.

3.2.2.2 Content or context validity

As the name suggests, content validity refers to the extent to which the instrument measures the full construct domain and whether the items included in the instrument are relevant for the purpose and context of the assessment (Fulcher 1999: 226). Messick (1989: 5) states that content validity “is evaluated by showing how well the content of the test samples the class of situations or subject matter about which conclusions are to be drawn”.

Fulcher (1999: 227) argues that content validity should include the level of test item difficulty, the quality of the rubrics and the accuracy of the scoring key. It follows that construct irrelevance variance results from inaccurate rubrics and scoring instruments as well as from items that are too easy or difficult for the target group. Aspects of content validity, discussed by theorists (Alderson *et al.* 1995: 176; McNamara 1996: 96; Brualdi 1999: 3; Fulcher 1999: 492), include

- the extent to which the items in the instrument measure the full construct domain;
- the relevance of the tasks to the purpose and context of the assessment;
- the construction of the instrument according to specifications related to the ability (construct) being tested;
- the extent to which the test or assessment is based on a theory of language ability measurement; and
- the level of task item difficulty.

Content validity depends on the adequacy of the content sampled in an assessment for the purpose of measuring a particular domain of knowledge, skill or trait. Procedures employed to obtain evidence of content validity involve:

- the compilation of a table of specifications to act as a framework;
- enumerating the information covered by the test;

- the number of items dealing with each content item;
- the manner in which these items are organised.

Sources of information include:

- syllabi;
- textbooks;
- teachers;
- curriculum developers;
- subject experts (to assess the content of the test and to rate each item).

Du Plessis and Weideman (2014: 131-132) point out that Weir (2005) “prefers to speak of context rather than content validity so as to reflect a socio-cognitive approach to language testing”. Weir (2005: 19) defines context validity as “the extent to which the choice of tasks in a test is representative of the larger universe of tasks of which the test is assumed to be a sample”. Weir (2005: 19) believes that there should be consideration of the “linguistic and interlocutor demands made by the task(s) as well as the conditions under which the task is performed”. This is particularly relevant to the challenging environment of the multicultural and multilingual target group of the current research and the difficulty is compounded by the distance-teaching mode of delivery.

3.2.2.3 Construct validity

The original definition of construct validity was that of Cronbach and Meehl (1955: 283), who describe a construct as “a postulated attribute of people, assumed to be reflected in test performances”. In the context of assessment, a construct is the specific ability, skill or aspect of a skill that test designers aim to measure by employing a specific instrument.”

Messick (1989: 5) defines constructs as “inferences about underlying processes or structures”, and Weideman (2006: 75) describes construct validity as “an analysis that indicates whether the theory or analytical definition (construct) that the test design is built upon is valid”. Kane and Bridgeman (2017: 514) describe the process of construct validation as “marshalling evidence in the form of theoretically relevant empirical

relations to support the inference that an observed response consistency has a particular meaning”.

Kane and Bridgeman (2017: 514) agree with Messick (1975: 955) who argues the centrality of construct validation in the assessment process, and states that “The process of construct validation... links a particular measure to a more general theoretical construct, usually an attribute or process or trait, that itself may be embedded in a more comprehensive theoretical network”. The views of Messick are seminal to the theories of validity and validation, and have been discussed and interrogated further in Section 3.3.1.

Kane and Bridgeman (2017: 501) also point out that:

In most assessment contexts, the question is not whether an assessment measures the trait or some alternate variable but rather the extent to which the assessment measures the trait of interest and is not overly influenced by sources of irrelevant variance.

Thus, an evaluation or measurement of irrelevant variance is essential in order to ensure that the assessment is valid (i.e. that it measures the “trait of interest”). As Kane and Bridgeman (2017: 501) go on to explain:

Messick (1975; 1989) made the evaluation of plausible sources of irrelevant variance a cornerstone of validation, and he made the evaluation of construct-irrelevant variance and construct under-representation central concerns in his unified model of validity.

In the light of the foregoing, it would seem that while the concept of construct validity might seem simple at first glance, further examination reveals its complexity. According to Fulcher and Davidson (2007: 7), it is difficult to define construct validity because of the difficulty in defining the term “construct”. Fulcher and Davidson (2007: 7) point out that a construct does not constitute a physical ability, but refers to a psychological trait. Examples of these include intelligence, achievement, motivation, anxiety, attitude, dominance, and reading comprehension (Ebel & Frisbie 1991: 108). These are

theoretical conceptualisations of areas of human behaviour that cannot be measured or observed directly.

In the context of language testing, Weideman (2006: 75) reiterates that it is extremely difficult to define a particular construct. Weideman (2006: 75) notes that a “blueprint for... a particular language ability cannot simply be plucked out of thin air. It has to be theoretically stable and robust, and stand up to the test of being empirically validated as a construct, as well as to the scrutiny of experts”. Du Plessis and Weideman (2014: 131) expand on this idea by observing that “Construct validity is achieved when the abilities to be assessed are founded on accepted theories of language, cognition and communicative competence.”

According to Patterson and Weideman (2013: 108), the construct should be clearly articulated based on a “theoretically defensible definition of what it is that should be measured”. As Du Plessis and Weideman (2014: 131) point out, this implies a strong correlation between what the assessment claims to measure and “indices of behaviour that one might theoretically expect it to correlate with” (Weir 2005: 18).

Thus, construct validity constitutes the extent to which a test or performance measures an underlying psychological construct or structure, such as language ability (Brualdi 1999: 2). Bachman (1990) posits that a construct is a way of defining ability and provides a means of theorising about its relationship to other abilities as well as to observed behaviour. It should be acknowledged, however, that it is extremely difficult, if not impossible, to measure an ability with complete accuracy. At best, as close an approximation as possible can be achieved, based on a relevant observable behaviour. Therein lies the problem of construct measurement. This is acknowledged by Messick (1989: 13), who points out that validity is a matter of degree and states that “what is to be validated is not the test or observation device as such, but the inferences derived from test scores or other indicators”. According to Rambiritch (2013: 113), Messick (1989) envisages validity as an “inductive summary of both the existing evidence for and the potential consequences of score interpretation and use”.

Fulcher and Davidson (2007: 7) explain that concepts become constructs when they can be linked to a test of some kind that will result in an observable outcome. In other

words, concepts should be defined in such a manner that they become operational. It is important also to establish the construct in the context of a theory that relates it to other constructs. In language assessment, constructs such as reading and writing ability are “latent traits” (Hattingh 2009: 22) which can be measured only indirectly by observation of behaviour elicited by appropriate testing and assessment (Henning 1991: 183).

Construct validity is thus linked to cognitive ability which Hattingh (2009: 92) describes as “the degree to which test tasks activate the same cognitive processes as writing in a real-life context”. As Hattingh (2009: 92) points out, test-takers will use “resources such as their content knowledge, which may be existing background knowledge or provided by the task input” in order to respond to the question asked. As has been discussed in Chapter 2, in connection with the target group of the present study, this might be problematic in this context because of the diversity of the cultural and linguistic backgrounds of the group although, on the other hand, the input provided by the materials is shared by all. The challenge lies in making the material generally accessible despite these differences. The same would apply to the rating scale in order for it to serve a formative function.

Construct validity is also concerned with the extent to which the test or measurement is grounded in the theory of language ability and assessment. This theoretical basis should be operationalised by means of clear definitions and measurable, reliable indicators (Garson 2006: 2). The proposed construct should also correlate with existing theory based on related studies that use other measures. Messick (1989: 13) stresses that validation of a construct should thus be the process of gathering evidence to support the contention that the assessment measures the construct that it is intended to measure, and that the scores provide an accurate reflection of this underlying ability or trait. As discussed in Chapters 2 and 6, it would appear, for example, that there is some construct under-representation because the rating scale is too generalised and does not make provision for some of the criteria as they are stated in the outcomes of the module (Appendix C). It is this issue that prompted the present research.

Further validation of construct validity can be undertaken by means of frequent use and testing of the construct in various settings. As Hattingh (2009: 21) reminds us, the goal of validation is to “determine the meaning of scores from the test, to assure that the

scores mean what we expect them to mean”. Hattingh (2009: 21) points out that “the more a construct is used by researchers in more settings with outcomes consistent with theory, the more its construct validity”. This is because a good construct is characterised by a sound theoretical basis, operationalised by clear definitions of measurable indicators (Garson 2006: 2). On the other hand, a poor construct might be at odds with related theory, or demonstrate flawed operationalisation to the extent that indicators can be interpreted in different ways by different researchers. As will be discussed further in Section 3.3, the concept of construct remains elusive and the debate on the position of construct validity in relation to other validity types remains unresolved.

3.3 MOVING TOWARDS A UNIFIED INTERPRETATION OF VALIDITY

As early as 1957, the traditional view of validity was questioned. Loevinger (1957) criticised the categories of validity (content, predictive, concurrent and criterion-related validity and construct) for not being sufficiently distinct or carrying equal weight. Loevinger (1957) argued that the parts represent options, rather than components, of validity. Loevinger (1957) suggested that content and criterion-related validity serve as supporting evidence for construct validity, rather than as separate types of validity. Loevinger (1957) believed that only construct validity provided a scientific basis for establishing the validity of an assessment instrument.

During the last three decades of the twentieth century, the traditional concept and scope of validity was interrogated increasingly. For instance, Guion (1980) criticised the trinitarian approach to validity, which comprises criterion, content and construct validity. Landy (1986) compared traditional validity processes with stamp collecting, to emphasise the lack of unity between the different types of validity.

The movement towards a unified approach was indicated by the American Educational Research Association, American Psychological Association and National Council on Measurement in Education Standards (AERA, APA & NCME, 1985: 9), which described validity according to the traditional categories, but pointed out that these categories should not be seen as separate types of validity. This move towards unity was

echoed at a much later stage by Douglas (2000: 257 - 258), who illustrates the complex nature of validity by comparing it to a mosaic in that “each piece of ceramic or glass is different ... from each other piece, but when they are assembled carefully... they make a coherent picture which viewers can interpret”. For this reason, Douglas (2000: 257 - 258) coined the term “validity mosaic” in order to “characterize the process”.

3.3.1 Messick’s unitary approach to validity

In his seminal work on validity, Messick (1989) posits an even more unified approach than that implied by the mosaic metaphor of Douglas (2000: 257 - 258). Messick (1989: 13) emphasises the central role of construct validity, which he describes as “the integrating force that unifies validity issues into a unitary concept”. This is because “it binds the validity of test use to the evidential basis of test interpretation”. Messick (1989: 13) sees validity as “an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment”. According to Messick (1989), a unified view of validity, with construct validity at its core, is the solution to the problem of fragmentation. Messick’s model is depicted in Table 3.1.

Table 3.1: Facets of test validity

	Test Interpretation	Test Use
Evidential Basis	Construct Validity	Construct Validity + Relevance/ Utility
Consequential Basis	Value Implications	Social Consequences

Source: adapted from Messick (1989:20)

This matrix portrays validity as a unified model with construct validity as a central, unifying component. Content and criterion validity are seen as aspects of construct validity. The matrix also presents a progressive classification of validity, which consists of the source of justification of the testing, and includes consideration of evidence and consequence, and the outcome or function of the testing, including test interpretation and/or use. The matrix graphically demonstrates Messick’s contention that construct validity is the unifying and integrating force that “binds the validity of test use to the validity of test interpretation” and also “binds social consequences of testing to the

evidential basis of test interpretation and use” (Messick 1989: 10). In other words, according to Messick’s model, constructs provide “the structure for validation” as well as “the glue” that binds all the elements (Kane & Bridgeman 2017: 519).

Kane and Bridgeman (2017: 523) point out that, while Messick considers “construct underrepresentation and construct-irrelevant variance as serious threats to validity in all cases”, he believes them to be “especially serious if they lead to adverse consequences”. As Messick (1989: 42) avers: “This is precisely why unanticipated consequences constitute an important form of validity evidence. Unanticipated consequences signal that we may have been incomplete or off-target in test development and, hence, in test interpretation and use”.

Messick’s interpretation of validity is supported by Weir (2005) and Shaw and Weir (2007: 3), who consider construct validity as central to the interaction between cognitive ability, context validity and scoring validity. The main elements of Weir’s (2005) framework comprise context validity, theory-based validity, scoring validity, consequential validity and criterion-related validity, depicted within a unified model, with construct validity as a super-ordinate category uniting the various elements.

Weir’s unified concept of validity offers a useful guideline to the validation process and his framework has been discussed in more detail in Chapter 4, which includes this process. What is of relevance to the current discussion is that Weir (2007) stresses that latent constructs are being assessed in language testing and that, besides cognitive constructs, social aspects (such as context and audience) should be considered when assessing students’ abilities. Weir (2005: 47) believes that:

The more comprehensive the approach to validation, the more evidence collected on each of the components of this framework, the more secure we can be in our claims for the validity of a test. The higher the stakes of the test the stricter the demands we might make in respect of all of these.

In an overview of Messick’s influence, Kane and Bridgeman (2017: 522) enumerate the following “basic conclusions” consistently emphasised by Messick. These are:

- Validity is a unified concept.
- Construct validity is the “framework for the unified model of validity” (Kane & Bridgeman 2017: 522).
- Validation is a type of scientific inquiry, not a checklist.
- Validity involves values.
- Validity appraises the social consequences (both intended and unintended) of score use.

In summary, Messick’s theory has had a profound effect on the current view of validity, despite adjustments such as those of Weir (2005) and Shaw and Weir (2007). As Hattingh (2009: 232) points out, “modern interpretations of validity regard it as a unified concept comprising different types of validity. Construct validity is generally accepted as an overarching term, and regarded as a function of the interaction between various types of validities”. This is a result of the widespread influence of Messick’s arguments.

3.3.2 Critique of Messick’s theory

Although the modern concept of validity has been shaped by Messick’s theory, his views have not been accepted without question (Shepard 1993; McNamara & Roever 2006; Rambirtich 2013; Ryan 2014). Firstly, the unified model has been criticised for being too complex and therefore difficult to operationalise. Secondly, his unitary interpretation has been questioned. A third concern is whether test consequences should be included in the interpretation of validity. Finally, Messick’s emphasis on scoring as a determining factor of validity has been interrogated.

The chief criticism of Messick’s progressive matrix is that his unified model is too complex and thus impractical. Messick’s matrix has been described as opaque, “incomprehensible” and “demanding” (Shepard 1993: 429). Although Shepard is in broad agreement with Messick’s unitary interpretation of validity, he is of the opinion that Messick’s four-fold matrix can lead to a segmented view of validity. This is ironic since it is the very interpretation that Messick wishes to prevent. Shepard expresses concerns that the model may give the mistaken impression that the values can be

regarded as separate from the scientific evaluation of scores. According to Shepard (1993), the two rows in Messick's table could be misinterpreted as implying that issues of scientific validity should be resolved before addressing value implications. However, it should be remembered that Messick (1989: 62) does in fact stress that scientific and value issues should be dealt with simultaneously. He states that "scientific observations are theory-laden and theories are value-laden".

Secondly, Shepard (1993: 427-428) criticises Messick's interpretation of construct validity as presented in his matrix. Shepard argues that it is unclear whether the term "construct validity" refers to the part or the whole. Shepard (1993: 428) continues by observing that the matrix could imply a narrow definition of construct validity as score meaning (first cell only). Shepard (1993: 428) points out that construct validity should be linked to all the criteria implied by all four cells, not merely that of score meaning. It should be noted that Messick (1989) acknowledges that the boundaries of his matrix are not watertight, describing them as "fuzzy".

A third criticism raised by Shepard (1993: 429) is that the complexity of Messick's analysis makes it difficult to apply to specific test situations. Shepard argues that the sequential diagram might be confusing and could result in mis-identification of the criteria to be addressed in designing a relevant test or assessment. For this reason, Shepard appeals for a more straightforward interpretation.

McNamara and Roever (2006: 800) also suggest a re-interpretation of Messick's matrix. This re-interpretation includes:

- What test scores are assumed to mean.
- When tests are used.
- The use of evidence in support of claims (to ensure fairness).
- The reasoning and empirical evidence supporting the claims made about candidates based on their test performance.
- Whether these interpretations are meaningful, useful and fair in the specific context.
- The overt social context of testing.

- The social values and assumptions underlying test constructs and the sense we make of test scores.
- What occurs in our education systems and the larger social context when we use tests?

Table 3.2: Interpretation of Messick's validity matrix

	What test scores are supposed to mean	When tests are actually used
Using evidence in support of claims: test fairness	What reasoning and empirical evidence support the claims we wish to make about candidates based on their test performance?	Are these interpretations meaningful, useful and fair in particular contexts?
The overt social context of testing	What social and cultural values and assumptions underlie test constructs and the sense we make of test scores?	What happens in our education systems and the larger social context when we use tests?

Source: McNamara & Roever's (2006: 800)

As Ryan (2014: 5) notes, McNamara and Roever's belief that if a language test has positive psychometrical qualities, it does not necessarily mean that it will have positive social consequences. In support of the theories of Shohamy (2005), Ryan (2014: 5) notes that "fairness and ethics in language testing" requires researchers to describe "the nature of test consequences" as well as "the challenges and complexities that researchers may encounter" (Ryan 2014: 5). Furthermore, Shohamy (2005: 49) points out that consequences often occur "outside the domains which the researcher examines" and hence are not immediately noticeable. This is a challenge to studies like the current one, and it can be argued that the best that the researcher can achieve is to attempt to mitigate the most obviously negative consequences in the present fluid situation. This is addressed in the final chapters of the thesis.

The concern about the complexity of Messick's matrix is reiterated by more recent researchers such as Rambiritch (2013: 116), who refers to "the daunting task of unravelling Messick's concept of validity" and the problem of its accessibility. Rambiritch (2013: 116) believes that:

While Messick has made an influential contribution to the field of testing, his work is not easily accessible to the lay person who needs to understand the field of testing, nor does he present us with a framework or guidelines to assist in the designing of tests that are accessible and transparent.

A similar viewpoint is expressed by Van der Walt (2012: 1), who acknowledges that “The unitary concept of validity, propagated by Messick (1989), has been very influential and informs many professional standards and codes for assessment”, but believes that it has “not provided clear guidelines for the validation of tests and is not easy to implement in practice”. This, according to Van der Walt (2012: 1) has given rise to attempts by researchers, such as Kane (2006) and Weideman (2009; 2012), to obtain “conceptual clarity”.

A further question is whether a unitary concept of validity is necessary. Borsboom *et al.* (2004: 1069) state that they “do not see the need for a unified validity concept... because we think there is nothing to unify”. They argue that, although a case can be made for an over-arching term to unite the various types of validities, the use of construct validity as a unifying term has caused the meaning of the concept to become unclear. They believe that more effort should be made to investigate the semantics of validity and thus answer the fundamental question of what makes a test valid.

Weideman (2009; 2012) also questions the need for a unitary concept of validity and validation, in this case, in the context of language testing. Weideman (2009: 2) believes that the reason for attributing this unifying function to construct validity is unclear. In addition, Weideman (2009: 2) points out that Messick’s (1980: 1025) reference to a test as “accomplishing its intended purpose” could lead to unhelpful, round-about references, since it is unclear why a test would not be valid if it does what it is supposed to do, namely, to achieve the intended result of producing the required measurement. Weideman (2009:10) argues further that a test’s “results could become the evidence or cause for certain desired (intended or purported) effects”. He suggests an alternative matrix in an attempt to solve this impasse (Weideman 2009: 240; 2012: 6).

Table 3.3: The relationship of a selection of fundamental considerations in language testing

Inferences made from test scores	adequacy of ...	appropriateness of...
	depends on multiple sources of empirical evidence	relates to impact considerations/ consequences of tests
The design decisions derived from the interpretation of empirical evidence	is reflected in the usefulness/utility or (domain) relevance of the test	will enhance and anticipate the social justification and political defensibility of using the test

Source: (Weideman, 2009: 240)

Weideman (2012: 10) claims that this “representation still follows Messick’s argument but, rather than validity, articulates the coherence of a number of assessment concepts”. Weideman (2012: 10) believes that concepts such as the technical adequacy and appropriateness of assessment instruments, the “technical meaningfulness” (interpretation) of measurements, as well as “their utility, social impact and public defensibility” will result in a useful reconceptualisation of “not only validation and validity, but all of our efforts at designing assessments responsibly”.

Weideman (2009: 6; 2012: 10) stresses that the adequacy and appropriateness of inferences arising from test scores depends on “multiple sources of empirical evidence”. Weideman (2009: 6) also emphasises the importance of the relationship between the fundamental considerations in language testing, and highlights the impact of the consequences of test design (derived from interpretation of empirical evidence obtained from the test scores), particularly how this reflects the “usefulness/utility or (domain) relevance” of the test and also the degree to which it “will enhance and anticipate the social justification and political defensibility of using the test”.

In essence, Weideman (2012: 11) feels that “we should seriously consider abandoning the notion of an overarching validity in favour of... an idea of responsible test design”. This responsibility would include consideration of social consequences and values. In the South African context, Weideman (2012: 11) states that the:

Current debates in South Africa about standardisation and equivalence can be deepened if we examine ways of going beyond conventional notions of validation

and validity, and take responsible design criteria to constitute the overriding condition(s) for the development of assessment instruments.

Johnson *et al.* (2015: 128) adopt a critical realist approach, “seeking to discern some of the structures and generative mechanisms that help to explain why an assessment system operates the way it does”. This approach is based on “an assumption that the world comprises socially patterned behaviours” and that research is able to “gather observable evidence about the nature of these patterns” (Johnson *et al.* 2015: 128). In addition, the social world is seen as a “stratified open system”, which, according to Johnson *et al.* (2015: 128), indicates that the “the things observed can be indicative of multiple, interacting events which themselves are not directly observable”. The implications for research design include an understanding of “multiple interacting elements” that “contribute to an observation outcome”. It thus follows that multiple research methods will need to be employed to “capture some of the complex, interacting elements” (Johnson *et al.* 2015: 129). There is, however, still a danger “for designed social actions, such as education policies, to lead to unintended consequences” (Johnson *et al.* 2015: 129). In the South African context, for instance, it can be argued that the seemingly ill-advised and sudden attempt to introduce outcomes-based education (OBE) was unsuitable to the social context and that the subsequent, misdirected focus on curriculum changes (combined with multiple socio-economic and socio-political factors) might have led to the unforeseen (and undesired) consequence of producing students who are unprepared for tertiary education.

Other critics of Messick’s unified model (Crocker 2003; Lissitz & Samuelsen 2007) propose a return to content validity in preference to construct validity as a central aspect of validation. Lissitz and Samuelsen (2007) go as far as to suggest a removal of the term construct validity when discussing test meaning. This is, however, contested by Gorin (2007), who considers that the changes suggested by Lissitz and Samuelsen (2007) would signal a return to the traditional, problematic validity theories and measurement practices, which Gorin (2007: 457) describes as having been “tried and discarded”. While agreeing with Lissitz and Samuelsen (2007) that there should be more emphasis on internal validity evidence, Gorin (2007: 457) argues that “Constructs exist across all assessment contexts” and should not be limited to one type of validity, such as content validity. Gorin (2007: 457) disagrees that content is sufficient to assess the validity of a

test, preferring construct validity as a unifying factor. Gorin (2007: 457) also avers that “what Lissitz and Samuelsen present as a radical change in validity terminology is more appropriately characterized as an issue of semantics or perhaps terminological preference”. Other researchers (Borsboom *et al.* 2003; Moss 2007; Weideman 2009), while not agreeing with all of Messick’s ideas, also disagree with the idea of returning to content validity as the paramount aspect of validity.

It would seem that one of the main reasons for difficulties experienced in the application of Messick’s theory is the amount of available evidence generated by different kinds of validity and the lack of clarity regarding the prioritising of this plethora of evidence. This leads to the criticism that construct-centred validation is an unreachable goal (Kane 1992; 2004; 2017; Shepard 1993; Lissitz & Samuelsen 2007) and that, in practice, test assessors frequently resort to providing only partial evidence.

Furthermore, as Borsboom *et al.* (2004: 1061) and Lissitz and Samuelsen (2007: 437) point out, theorists and test practitioners frequently do not appear to have the same concepts in mind when investigating the validity of a test. According to Shepard (1993: 429), this is because standards are not organised or prioritised in a coherent conceptual framework. These criticisms indicate the need for a reformulation and simplification of Messick’s matrix rather than an outright rejection of it (Shepard 1993: 429; Kane 2004: 136). As Kane (2004: 140) points out, “Construct validity has been useful as a unifying framework on a theoretical level, but has not, in itself, been an effective unifying influence on an operational level”.

Another concern about Messick’s interpretation is the question of whether the consequences of tests should be incorporated into the concept of validity and, if so, how this can be achieved (Weideman 2009; 2012). As discussed in Section 3.3.1, in a significant development, Messick (1989; 1996) introduced a social dimension of validity by arguing that the traditional viewpoint ignores the social implications of scores and the consequences of the decisions that are made based on these scores. Messick thus added a further dimension to the concept of construct validity by arguing that it also “binds social consequences of testing to the evidential basis of test interpretation and use” (1989: 10). Messick’s opinion (1989: 18) is that if “questions are whether the potential and actual social consequences of test interpretations and use are

not only supportive of the intended testing purposes, but at the same time are consistent with other social values... social values and social consequences cannot be ignored in consideration of validity".

These ideas have had an influence on recent theories and practice, and have given rise to the current consideration of the social consequences of a test and whether the results will be harmful or beneficial to the test-takers. For instance, Bachman and Palmer (1996: 30) amplify these views by stating that “the very acts of administering and taking a test imply values and goals and these have consequences. Similarly, the uses we make of test scores imply values and goals and these have consequences”. Bachman and Palmer (1996) note that testing has consequences for all stakeholders in the testing process – not only for test-takers, but also for teachers and educational systems. Consequential validity can thus be measured by the degree to which decisions taken on the basis of results or scores promote the well-being of those who will be affected by these decisions (McNamara & Roever 2006; Ryan, 2014: 4 - 5).

The concept of consequential validity as posited by Messick has particular relevance to the area of language testing. McNamara and Roever (2006: 32) point out that “language is rooted in social life and nowhere is this more apparent than in the ways in which knowledge of language is assessed”. Weideman (2006: 72) signals the changing perspective of applied linguistics in the following comment:

The designed solutions to language problems that are the stock-in-trade of applied linguistics affect the lives of growing numbers of people. By calling for these designs to be accountable, applied linguistics has, in its most recent, postmodern form, added an ethical dimension that is lacking in earlier work.

In stressing the importance of emphasising consequences in the assessment process, Fung (2017: 101) points out that assessing “students’ learning in higher education is a high-stakes activity” and adds that “As well as being extremely time-consuming, both for students and for assessors, assessment can determine students’ futures”. According to Fung (2017: 101), educators can shape students’ orientation to their studies by means of careful assignment design. Fung (2017: 101) adds that in assigning scores, assessors might be influencing students’ future access to further study or to a profession, and

could also influence “students’ self-confidence and self-concept”. This has been discussed in Chapter 2 with reference to the target group and the challenging context of distance learning in which assignment feedback is the main (and sometimes only), communication between lecturer (or tutor) and student, and thus assumes central importance in formative assessment. It is also a challenge to articulate the criteria in a way that is clear and unambiguous, given the socio-cultural and multilingual composition of the group.

Rambiritch (2013: 116) acknowledges that the issue of the social dimension introduced in Messick’s model is possibly the reason for Messick’s theory gaining such widespread acceptance, since it signals a necessary change in the focus of testing which, until then, had placed a heavy emphasis on psychometrics. However, despite this shift in focus, Rambiritch (2013: 116) points out that Messick’s theories still rely heavily on empirical data and statistical relevance.

Similarly, Shepard (1993: 427) voices the concern that the “very issues” highlighted by Messick could risk being ignored by researchers because the “categories of use and consequence appear to be tacked on to ‘scientific validity’, which remains sequestered in the first cell”. Shepard cautions that the separate cells of Messick’s model could give the impression that the researcher or tester should first pay attention to the “scientific question of test score meaning and then proceed to consider value issues”. This concern is also mentioned by McNamara and Roever (2006: 248), who state that “despite Messick’s efforts to build a unitary approach to validity that acknowledged the social meaning of tests, validity theory has remained an inadequate conceptual source for understanding the social consequences of tests”.

On the other hand, Alderson and Banerjee (2002) take issue with the implication that test developers should be held responsible for the use or misuse of tests and their results. They question whether the term “consequential validity” is a genuine concern or merely a “political posture” (Alderson & Banerjee 2002: 79). Lissitz and Samuelson (2007: 445) adopt a more moderate approach, arguing that, although it is sometimes necessary to consider the impact of a test on its stakeholders, any unintended or unwanted effect that the test might have “should not be considered relevant to the question of whether the test is valid”.

For the purposes of the current study, the importance of the social consequences of assessments is acknowledged, particularly in the South African context where the majority of students do not claim English as a home language, and many come from disadvantaged socio-economic backgrounds. Examination scores are used to make decisions that have serious and far-reaching consequences for students' academic future, and can ultimately impact on their career prospects. It is thus essential that examination and assignment marks should be fair and accurate, and should be interpreted correctly. The importance of interpretation is emphasised by Weir (2005: 12) who sees it as a component of validity:

Validity is perhaps better defined as the extent to which a test can be shown to produce data, i.e. test scores, which are an accurate representation of a candidate's level of language knowledge or skills. In this revision, validity resides in the scores on a particular administration of a test rather than in the test *per se*.

However, it should be borne in mind that, in order to produce results that give an accurate reflection of the students' abilities, one needs an appropriate test and an accurate rating instrument that measures the construct to be tested. In the context of measuring scales, Fulcher and Davidson (2007: 16) refer to the importance of the fairness governing the assignment of scores and the importance of measurements being regarded as an accurate reflection of the ability or trait in question. The danger of disregarding consequences and context has been pointed out by Alderson *et al.* (1995: 170), who express concern that measuring instruments are frequently used for assessing abilities other than those for which they were originally designed.

According to Kane and Bridgeman (2017: 505) "the terms fairness and bias can be interpreted as covering roughly the same ground, with fairness being defined as the absence of bias". However, fairness can include a broader range of issues (including those concerned with social equity), implying that "a fair test is comparable from person to person and group to group" and thus "impartial and lacking in prejudice or favouritism" (Kane & Bridgeman 2017: 207). In contrast, the interpretation of bias may be "narrower and more technical" (Kane & Bridgeman 2017: 207), often "akin to the

notion of bias in the estimation of a statistical parameter” (Kane & Bridgeman 2017: 207).

In the present study, the importance of considering the social consequences of tests was recognised, as advanced by Messick (1989; 1992), and careful test and instrument design that bears in mind the purpose of the test and the construct to be assessed was emphasised. This led to the selection of relevant test content and criteria being included in the rating instrument. Thus, fairness and lack of bias are essential for an assessment to be valid, although fairness covers the broader field of social consequences.

Despite the criticism of Messick’s theory, the positive effects of his unified concept of validity must be acknowledged. For example, Moss (2007) points out the value of a unified approach to validity, praising it for the guidance it gives in investigating validity. Moss (2007: 474) is of the opinion that Messick’s framework assists the researcher in progressing from conceptualisation to test development and finally implementation. Gorin (2007) agrees that the unitary model of validity is valuable because it provides a flexible vocabulary and thus encourages cross-disciplinary discussion and implementation. A similar opinion is advanced by Kane and Bridgeman (2017: 522), who acknowledge the value of Messick’s approach in providing “a comprehensive framework for validation”, but remind the reader that this is “a framework intended to encourage and guide conversation and investigation. It was not intended as an algorithm or a checklist for validation”.

In addition, Messick’s ideas have inspired further research that has resulted in the simplification and streamlining of his matrix, enhancing its accessibility (Bachman 1990; Weir 2005; Shaw & Weir 2007). Furthermore, the introduction of the aspect of consequential validity has broadened the validity concept from a mere reflection of test scores to a consideration of the purpose and impact of tests. This has resulted in increased awareness of the consequences of scores and thus the importance of an accurate and fair test and rating instrument (Weideman 2012; Ryan 2014: 3; Kane & Bridgeman 2017: 505).

Debates on validity theory are ongoing and likely to remain so for a considerable time, but each one contributes to progress in this complex field. Kane and Bridgeman (2017: 522) comment on this issue as follows:

As is true in most areas of scientific endeavor, theory development is an ongoing dialogue between conjectures and data, between abstract principles and applications, and between scholars with evolving points of view.

3.4 VALIDITY AND RELIABILITY

Reliability or scoring validity refers to the consistency and credibility of tests scores. Anastasi (1976: 103) defines reliability as “the consistency of scores obtained by the same persons when re-examined with the same test on different occasions or with different sets of equivalent items, or under other variable examining conditions”. This description has been reiterated by later researchers. For example, Henning (1991: 285) describes reliability as "the capacity of the assessment procedures to rank-order the same samples of writing performance consistently in the same way". Jones (2001: 1) believes that a test is reliable if it can produce similar results on different occasions, and Weir (2005: 23) is of the opinion that the measure of reliability is the extent to which unbiased, stable and consistent test results are produced in a particular situation, and from one situation to the next. Potential sources of unreliability include:

- time of day;
- students’ background;
- motivation;
- state of health;
- degree of fatigue;
- rater characteristics;
- assessment scales;
- scoring procedures.

These factors are not a reflection of the student’s innate ability and for this reason they interfere with the accuracy of test results. Bachman (1990: 163 - 166) also discusses the

influence of aspects such as communicative language ability, test method, test-takers' personal attributes and other random factors. Bachman (1990: 161) points out that if the interference of such aspects can be minimised, the test will give a more accurate indication of the tested ability, and thus reliability will be increased. As Hughes (1989: 36) notes, it is not possible to obtain a reliable test score from an unreliable test.

As noted by researchers (Bachman 1990; Weir 2005), important factors influencing reliability include rater bias and the characteristics of rating scales. Although rater bias undoubtedly has a negative impact on reliability, Lane (1999: 6) makes the point that investigating rater consistency will be useful only if test content reflects the specified construct and if rating scales contain relevant criteria, and do not include those that are irrelevant to the construct.

A further concern is the influence that scoring methods have on scores. Different scoring methods (for example, holistic or analytic scoring) focus on different characteristics of test performance, and will thus lead to the assigning of different meanings to scores (Lane 1999: 6). These factors are discussed in more depth in the chapter on validation (Chapter 4).

The relationship between reliability and validity has been problematic. Weideman (2006: 74) clarifies the difference between the two by stating that validity “normally refers in this context to the power of a test to assess what it is designed to do, and reliability to the consistency with which it measures”. As Hattingh (2009: 36–37) points out, “a test must be proved reliable in order to establish other empirical types of validities”. However, she adds that a “reliable test is not necessarily valid, as it may give consistent results, but may not be measuring what it claims to, or may not be appropriate and meaningful within the context in which it is used” (Hattingh 2009: 36–37).

Both validity and reliability are essential aspects of assessment, yet the tension between them has often led to a trade-off, frequently with an emphasis on reliability at the expense of validity. According to traditional theory, reliability and validity counter-balanced each other. Many researchers believed that the reliability of an assessment would need to be reduced in order to increase the validity (Alderson *et al.* 1995: 42).

Furthermore, theorists argue that the differences between the two concepts are unclear and thus lead to confusion (Marcoulides 2004: 183). This, in turn, results in the neglect of validity in favour of an emphasis on reliability (Huot 1990: 202). This opinion is reflected in Rozeboom's (1966:375) description of reliability as "the poor man's validity".

The emphasis on reliability was criticised as early as 1970 by McColly (1970: 149), who states that:

It is often said that reliability is the more important of the two... But really the inverse relationship is true. Be that as it may. The scholarly literature that deals with writing tests shows more apparent interest with reliability than with validity.

However, as Hattingh (2009: 37) points out, traditional assumptions about the tension between reliability and validity have persisted despite arguments in favour of validity as the central attribute, and the questioning of the belief that reliability is in itself a sufficient condition for the validity of an assessment (Popham 1981). While it should be acknowledged that reliability remains an essential aspect of a valid test, merely proving that a test is reliable is not sufficient evidence of its validity. A reliable test is not necessarily valid, because although it might provide consistent results, it could fail to measure what it claims to assess, and might be inappropriate in the given context (Hattingh 2009: 36). On the other hand, reliability should be established in order to ascertain empirical validities such as concurrent, predictive and construct validity (Bachman 1990; Brown & Hudson 2002).

A solution to the seemingly difficult relationship between validity and reliability might be reached if the unified view posited by Messick (1989) is adopted, which suggests that reliability is an aspect of validity and that an increase in reliability should be regarded as evidence in favour of the overall validity of an assessment. Reliability and validity are seen as aspects of a unified approach, consistent with Messick's concept of a unified concept of validity, and in keeping with the trend towards this unified approach. As Weir (2005: 43) points out, the widespread acceptance of this viewpoint has ensured that reliability and validity are no longer polarised in the current understanding of validity. For instance, O'Sullivan (2006: 195) stresses that a model that

views validity and reliability as related and contributory aspects of a unified concept of validity would be of greater value and usefulness than one that considers the two aspects separately.

Messick's theory has signalled a shift from the traditional emphasis on reliability (at the cost of validity) towards the construct that is being investigated, combined with a fair assessment and scoring process. The emphasis thus falls on scoring validity, a concept closely linked to that of reliability. Scoring validity has been described as "external validity" (Alderson *et al.* 1995) and incorporates the procedures associated with the production and interpretation of scores. In agreement with Messick (1989), Lane (1999: 3) states that scoring validity is related to construct validity because the validity of score interpretations "is dependent on the fidelity of the construct that is measured by the test and the resulting test scores".

Weir (2005: 43) affirms the importance of scoring validity in locating "reliability more centrally in the validation process". Aspects of scoring validity mentioned by Weir (2005:47) include the reliability of scoring procedures, the ability of test raters to score tests consistently, and the accuracy of statistical elements associated with the scoring of tests. These include item analysis, internal consistency, error of measurement, and statistical measurement of marker reliability. All of these checks and balances enable test scores to give an accurate and consistent indication of the test taker's ability. In other words, scoring validity should reflect the reliability of a test.

Shaw and Weir (2007) list the aspects of scoring validity as:

- criteria/ rating scale;
- rater characteristics;
- rating process;
- rater training;
- rating conditions;
- post-examination adjustment;
- grading and awarding.

These and other factors of scoring validity will be discussed in greater detail in Chapter 4, which deals with practical issues of validation. What is relevant here is the blending of validity and reliability under the super-ordinate function of scoring validity, and the inclusion of both *a priori* (criteria/rating scale, rater training) and *a posteriori* (post examination adjustment and awarding) elements in scoring procedures.

In the current study, the view was adopted of validity and reliability as united and complementary elements concerned with providing a fair, trustworthy and consistent measurement of a student's ability. This implies that scores are fair and accurate (Bachman 1990: 23; Alderson et al. 1995: 23; Kane & Bridgeman 2017: 505) and that the effects of external factors on performance and scores are minimised (Bachman 1990). It is imperative for a measuring instrument to be reliable and free from errors of measurement. However, it has been demonstrated that validity depends on reliability, but that reliability is not sufficient to prove the overall validity of a test or instrument. This is because a reliable test might produce consistent results, but will not be valid if it does not measure what it claims to be measuring.

3.5 CONCLUSION

In this chapter, the complex nature of validity has been investigated and types of validity have been described. The development of the concept of validity has been traced from the traditional one that considers the types of validity as separate entities, to the unitary approach posited by Messick (1989). Various criticisms of Messick's concept have been discussed, in particular the difficulties arising from the perceived complexity of Messick's matrix, and whether construct validity is in fact a useful "umbrella" concept in determining validity. Alternative frameworks that attempt to simplify and streamline Messick's matrix were examined. These and other frameworks have been discussed in more detail in Chapter 4, which covers the process of validation. In the current study, the value of Messick's theory is acknowledged, but criticisms relating to its complexity are noted, and also the consequent difficulty that could arise from the interpretation and implementation of this theory. The adjustments presented in frameworks such as those of Weir (2005), and Shaw and Weir (2007), were attempts to

address these issues, and are considered to be useful in the validation process (Chapter 4).

Furthermore, consequential validity (introduced by Messick 1989), which resulted in the more recent emphasis on fair and accurate tests scores and interpretations, was discussed. It is believed that purpose and context should also be considered as factors affecting the validity of scores. This theory has been developed and in some cases criticised by researchers, but has resulted in the consideration of the consequences of the interpretation of scores in the academic and social environment, and is thus of relevance to the target group of the current study.

The changing perspectives of the previously problematic relationship between validity and reliability was traced. It was demonstrated that reliability is no longer seen as a separate element but as an aspect of validity, closely related to scoring validity. It can be argued that this theory eliminates the previously perceived conflict between these concepts. This is a positive development, in keeping with the modern unitary concept of validity.

The aim of this study was to provide fair, accurate and objective assessment of the written work of the target group. This assessment is of great importance for both formative and summative assessment, and can have an impact on the students' academic and socio-economic prospects. A unified view of validity is accepted as it presents various types of evidence under an over-arching concept, thus improving the accuracy of the result.

While acknowledging that validity is a matter of degree, it is believed that the unified model provides sufficient evidence to ensure a fair, objective assessment that will reflect the construct being tested. It should be borne in mind that "While agreed marking rubrics with specified assessment criteria help with the development of shared understandings, assessment is not an exact science and can never be entirely objective" (Funk 2014: 11), although this does not prevent the researcher from striving to achieve as fair and objective a result as possible.

CHAPTER 4: THE VALIDATION PROCESS

“Validation is simple in principle, but difficult in practice.” (Kane 2006: 15)

4.1 INTRODUCTION

In this chapter, definitions of the term “validation” are followed by a discussion of the validation process with particular reference to the evidence-based argumentative approach as a fair, relevant and effective means of validation.

The focus is then narrowed to a discussion of validation models and frameworks as a foundation for the development of an assessment instrument that meets the criterion of a “balanced scale that gives adequate feedback for both teachers and learners while being as practical as possible” (Hattingh 2009: 145). The number of assessment levels to be used was also determined by the empirical means of pre-testing and piloting, as recommended by Weigle (2002: 127).

Furthermore, factors that have an impact on scores are described. The factors include those directly related to learning, teaching and assessment (such as characteristics of test-takers, inter-rater variance, the washback effect, and avoidance of test bias) as well as administrative and physical factors, and the possible impact of the assessment on institutions and society. The rating scale employed for the target module is then discussed, and the chapter concludes with a brief overview of the issues covered in the chapter.

4.2 DEFINITIONS OF VALIDATION

In order to be considered valid, it is imperative that an assessment “tests what it purports to test; that it tests a property that exists and can be measured” (Van der Walt & Steyn 2007: 141). In other words, validity has to be demonstrated by a process of validation.

Weir (2005: 15) describes the validation process as “a form of evaluation where a variety of quantitative and qualitative methodologies... are used to generate evidence to support inferences from test scores”. This view supports an observation by Bachman (2004: 265) that:

...we must collect evidence supporting the construct validity of interpreting this score as an indicator of the individual's ability and consider the value implications of various labels we could attach to this interpretation of the particular theories... upon which these labels are based.

It can be said that validation is validity made “visible” in the context of the specific assessment situation, which includes the purpose of assessment and the assessment instrument (Bachman 1990; Alderson *et al.* 1995; Weir 2005; Hattingh 2009). With reference to rating scales, Hattingh (2009: 49-50) states that:

Validation is the process of proving that an assessment instrument measures what it claims to measure... that the instrument is relevant to the purpose and in the context of assessment, that [it]... produces scores that accurately reflect learner abilities, that the scores are interpreted accurately and inferences made fairly.

In short, the validation process entails collecting empirical data on a rating scale and using this information as the basis for a logical argument, in order to demonstrate that inferences arising from the test results are appropriate for the purpose of assessment, as well as for the particular target population. This can lead to a “consideration of consequential validity and by extension the role of formative assessment” (Brualdi 1999: 1). As Frederiksen and Collins (1989: 27) point out, a “systematically valid” assessment “induces in the education system curricular and instructional changes that foster the development of the cognitive skills that the test is designed to measure”. This signifies a relationship between the concept of consequential validity and the notion of washback (discussed further in Sub-section 4.6.1.3) because it can be argued that, if the assessment has a negative effect on the development of the abilities it claims to measure, the validity argument is weakened. This would result in social consequences discussed further in Section 4.3 and 4.6.3 of this chapter.

Du Plessis (2014), and Du Plessis and Weideman (2014: 129) emphasise the multi-faceted nature of the validation argument. Du Plessis (2014: 25) claims that the validation process refers to the “systematic presentation of... evidence as unity within a multiplicity of arguments”. Du Plessis and Weideman (2014: 130 - 131) develop this concept in contrast to the view of validation as “the process of collecting evidence in support of inferences of ability made on the basis of test or examination scores”.

4.3 THE VALIDATION PROCESS

Although researchers such as Weigle (2002), Weir (2005), Bachman (2005), Fulcher and Davidson (2007), and Chapelle (2012), emphasise the importance of validation, it has been claimed that many existing language assessment rating scales generate insufficient evidence to prove the validity of language assessment instruments. As Turner (2000: 556) observes:

...one often wonders how scales are developed. With the important role that rating scales play in performance evaluation, one would think that the literature would abound with descriptions and procedures for scale construction. But, as we quickly learn, this is not the case.

Kane (2004: 1) argues that, although there are many sophisticated validity theories, the methodology of validation processes is generally ineffective. Kane (2004: 1) believes that, in many cases, more evidence is provided for technical characteristics, such as the reliability of assessment programmes, than for their validity (as narrowly defined). While he acknowledges that the “basic principle of construct validity calling for the consideration of alternative interpretations offers some protection against opportunism”, he adds that “like many validation guidelines, this principle has been honoured more in the breach than in the observance.... Most validation research is performed by developers of the test, creating a natural confirmationist bias” (Kane 2004: 140). Furthermore, as Hattingh (2009: 49) notes, studies seem to be contradictory as a result of employing different criteria and types of evidence, and can even be opportunistic in their exclusive use of easily accessed supporting data (Turner 2000; Alderson & Banerjee 2002; Schilling 2004). This is an extreme development of the “confirmationist

bias" mentioned by Kane (2004: 140) and later confirmed by Chapelle (2012: 25-26), who remarks that: "Many professionals in language assessment are responsible for developing validity arguments in connection with a testing program, where they play an 'advocacy role'". As Chapelle (2012: 26) points out, this could result in their validity arguments demonstrating a "confirmationist bias" as described by Kane (2004: 140).

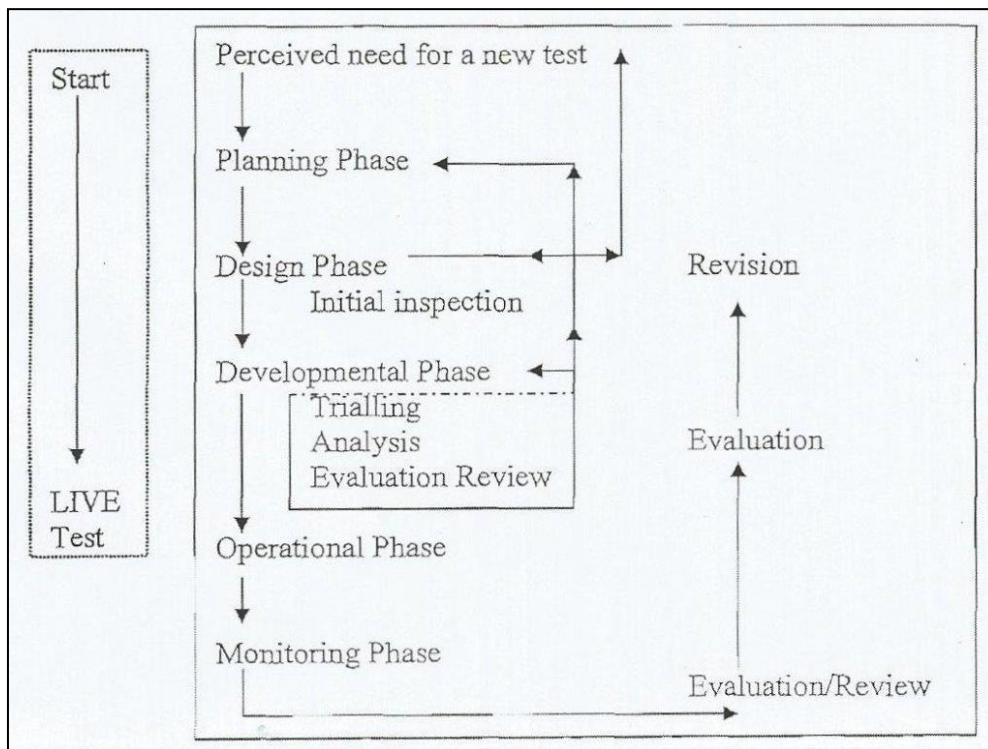
Chapelle (2012: 25-26) adds that this is particularly obvious in language testing in the business environment, where "some professionals have a confirmationist bias relative to the testing program they work with in addition to a refutationist bias toward competitors." According to Chapelle (2012: 25-26): "The ideal of objective appraisal that Kane sees as central to validation, in fact, may actually take place much less than one might assume in an environment where knowledge tends to be connected to experience, which is in turn connected to interest and even advocacy."

In a similar vein, Weir (2005: 15) points out that: "Most examinations lay claim to the numerous aspects of validity. However, what are often lacking are validation studies of actual tests that demonstrate this". An example of this problem is the theory of Messick (1989) in which the validity and the consequences of test use are addressed but which, as discussed in Chapter 3, has been criticised for the little guidance that has been provided regarding the practical and empirical investigation of these issues (Shepard 1993: 429; Weideman 2009: 2; Van der Walt 2012: 1; Rambiritch 2013: 116; Marshall 2016).

However, it could be argued that, in studies such as those of McNamara (1996), Saville (2001) and Taylor (2002), guidelines are provided on a general procedure for developing and revising assessment instruments. Research examining the validation of existing scales includes that of Messick (1992), Fulcher (1996), North and Schneider (1998), Lumley (2002), Weir (2005), Bachman (2005), Fulcher and Davidson (2007), Shaw and Weir (2007) and Hattingh (2009). In this research, the importance of clear, logical and explicit validation procedures based on a sound foundation of language theory is emphasised. To this end, McNamara (1996), Saville (2002) and Taylor (2002) agree that developing or reviewing an assessment scale includes three basic stages, namely: the design stage, the construction stage and a trial stage. In these studies, the outlines of validation procedure are provided, which can be developed further. The

guidelines provided were adopted also for the present research in formulating alternative scales for the target module (see Chapter 7).

Figure 4 shows Taylor's (2002: 2) graphic representation of the basic outline of a validation procedure. It illustrates a cyclical and iterative process. Each of the steps is made up of a series of validation activities to demonstrate how evidence is evaluated constantly "as an integrated set to determine to which degree the validity argument is supported" (Lane, 1999: 1 – 2).



Source: (Taylor, 2002: 2)

Figure 4.1: Model of the assessment instrument development process

Lane (1999: 1) believes that validation begins with "a construct in search of appropriate instruments and procedures", and evidence should be collected from the beginning of the design stage until after the administration of the assessment. Evidence from each stage offers new knowledge about the instrument, which is then included in its revision (Saville 2001: 5). The importance of *a posteriori* evidence is emphasised by McNamara (2000) and Saville (2001) who aver that evidence about the qualities and effectiveness of the assessment instrument can be collected only once the instrument is implemented. Fulcher and Davidson (2007: 21) go as far as to say that the validation process should

start at the final point (i.e. the consequences of testing), and then work backwards towards test design.

The assessment instrument should be revised, not only as a result of evidence indicating its validity for the particular assessment, but also with regard to any new issues that might arise in the development of the concept of validity, as described in Chapter 3 (Lane 1999; Douglas 2000; Bachman 2004). As Hattingh (2009: 51) observes: “The need for revision and validation of assessment instruments does not become satisfied once an instrument has been proven valid for one administration”. However, although Kane (2004: 151) agrees that validators should be constant critics of validity claims, he suggests that an instrument can be classified as valid for a particular assessment situation once the most problematic assumptions and inferences have been addressed and proven to be acceptable for the purpose and context. This applies to the current research study, during which the rating scales were subjected to thorough qualitative interrogation and quantitative analysis, resulting in an improved scale that is believed to address the “most problematical assumptions and inferences” (Kane 2004: 151) of the current scale and thus produces a rating scale more appropriate to the purpose as indicated in the module outcomes for ENG1501 (Appendix C) which served as the construct considered for this thesis. It is acknowledged that the proposed scale should be refined for future administrations, but this should not detract from its validity for the present situation.

Huot (1996: 161) adds a further dimension to the validation process by emphasising that local standards, including reading contexts and the background against which the design of assessments takes place should be considered when developing assessment procedures. Huot (1996: 161-162) believes that, in order to ascertain the role of context in a specific assessment, qualitative procedures such as interviews and observations should be employed to complement the role of quantitative validation procedures, and to prevent the scale from having unfair social consequences. This is one of the reasons why the present research project includes comments and questionnaires involving stakeholders, in addition to the quantitative data generated by statistical procedures. However, the heterogeneity of the target group presents a challenge, as one cannot refer to a single set of local standards in examining context.

Van der Walt and Steyn (2007: 141) agree with Huot (1996: 161 – 162) that validation is “dependent on the test results being used for the purpose for which the test is designed”. Furthermore, Van der Walt and Steyn (2007: 141) believe that, because various factors influence test score interpretation, sufficient evidence is essential to prove the validity of a test – in other words, “that it tests what it claims to test, and that it tests a property that exists and can be measured” (Van der Walt & Steyn 2007: 141). According to Van der Walt and Steyn (2007: 141), repeated use of a test can confirm and reinforce this validity. The validation “starts as a local affair, with repeated use of a test for one purpose only, and ultimately one can argue that validity becomes a property of the test”. Evidence should include “consideration of the purpose of the test, its content and method, intended (and possible unintended) consequences, potential decisions that can be made and the impact it may have on test-takers” and should “involve both descriptive- and decision-based interpretations that are made after relevant evidence has been collected” (Van der Walt & Steyn 2007: 169).

4.3.1 The argument - based approach to validation

Despite (and, possibly, in answer to) the criticism that there is incomplete guidance and research on validation procedures, Van der Walt (2012: 152) points out that a “number of frameworks for validation have recently been proposed” Linked to these frameworks is the development of an argument-based approach to validation, recommended by Cronbach (1988: 4) and reinforced by later researchers (Kane 1992; 2004; Kane, Crooks & Cohen, 1999; Bachman 2004; Fulcher & Davidson 2007; Hattingh 2009; Chapelle, 2012; Knoch & Elder 2013; Lydster & Brown 2017: 52). A validity argument entails providing sufficient evidence to evaluate the validity of an assessment, and to reach a conclusion based on the analysis of all evidence presented both in favour of and against the proposed interpretation of assessment results.

Bachman (1990; 2005) advocates an argument-based approach to validation, as do Touhmin (2003), Kane (1992; 2004) and Fulcher and Davidson (2007), in order to provide a systematic process applicable to a range of score interpretations and uses. Bachman (1990: 263-264) states that validation is “the process of building a case – articulating an argument and collecting evidence – in support of a particular

interpretation of test scores.” Bachman (1990: 263-264) describes this “evidence-centred interpretive argument” as having a dual function: firstly to provide guidelines (based on the test construct and specifications) for the design and development of the assessment instrument and, secondly, to provide a framework for collecting the “evidence necessary to support the intended interpretations and uses of scores”.

Bachman (2005: 14) concedes that, even if the score interpretations are valid, there remains a possibility that results could be used inappropriately for purposes other than those intended. He believes that sources of negative consequences (which he describes as “beyond invalidity”) must be included as far as possible in an argument that investigates the use of an assessment (Bachman 2005: 15-16). This could possibly occur, for example, if the scale were used for political ends, possibly in order to exclude members of a certain group. Taking issue with Messick (1994: 21), who avers that “the primary measurement concern regarding adverse consequences is that any negative impact on individuals or groups... should not derive from any source of test invalidity such as construct underrepresentation or construct-irrelevant variance”, Bachman (2005: 15) argues that “it is quite possible for adverse consequences and inappropriate uses of tests to occur that are not a result simply of sources of invalidity”. Bachman (2005: 15) explains that “it is possible for the results of assessments to be used inappropriately, even though these assessments are valid indicators of the abilities they are intended to measure”.

The argument-based approach to validation is advocated by Kane (1992; 2004), who is of the opinion that this approach serves to provide a methodology for validation (Kane 2004: 2). The systematic approach suggested by Kane (1992; 2004) is encapsulated in the six explicit steps that he distinguishes as constituting an argument for validity. These steps are:

1. Specify the intended score interpretations by means of stating a clear argument.
2. Evaluate the plausibility of the interpretive argument by examining its inferences and assumptions.
3. Adjust the argument, based on the evidence.
4. Examine the most problematic assumptions as identified in Step 3.
5. Evaluate the new argument generated as a result of Steps 3 and 4.

6. Identify potential weaknesses in the argument.

Hattingh (2009: 59) cautions that these steps “are not meant as a checklist, but serve to outline the argument-based approach in detail without being restrictive”. Furthermore, Hattingh (2009: 59) points out that “if any of these stages are omitted, the argument should justify such an omission”. Thus, the six steps provide clear and explicit guidelines, but also allow flexibility, depending on the purpose and context of assessment. Van der Walt (2012: 152) observes, “Kane’s argument-based approach is a major contribution to the discussion of the validation of language tests. It provides a systematic approach for the evaluation of a test at the development, trialing and implementation stages, and provides guidelines for a well-articulated validation methodology”.

Kane (2004: 139) uses construct validity as a starting point and unifying principle for the validation process, and emphasises the importance of operationalising the construct. As Du Plessis (2014: 2) points out, construct validity is central to the validation of any language assessment because, in the words of Messick, it “integrates considerations of content, criteria and consequences into a comprehensive framework for empirically testing rational hypotheses about score meaning and utility” (Messick 1995: 742). As Kane (2013: 118) points out:

...the evidence needed for validation depends on the inferences and assumptions inherent in the proposed interpretation/use, and these inferences and assumptions have to be specified in some way.... Validity theories face difficulties in identifying any particular kind of evidence as essential, or as irrelevant, because test-score interpretations and uses are so varied.... The argument-based approach gets around this problem by making validation requirements contingent on the claims being made.

Kane (2013: 18) adds that, in order to conduct this analysis, it is necessary “to implement two conceptually distinct steps”. According to Kane (2013: 18), these are to “state what is being claimed and... evaluate the plausibility of these claims”. It is the latter step that Kane associates with the validation argument.

According to Du Plessis and Weideman (2014: 129), the technical term, “validation”, has been coined by assessment experts to refer to the process of collecting evidence in support of inferences of ability made on the basis of test or examination scores (Kane 2004; Weir 2005; Bachman & Palmer 2010; Chapelle 2012; Van der Walt 2012). Weideman (2013: 13) develops this further by describing validation as a design principle requiring test developers “to systematically integrate multiple sets of evidence in arguing for the validity of a test” (Weideman 2013: 19). This was the aim of the current research in which a multi-pronged process was followed by combining quantitative evidence with a qualitative approach that considered the viewpoints of stakeholders at various stages of the process.

Du Plessis (2014: 2) agrees with Weideman (2012) that evidence should include:

- the selection of content;
- the context of assessment;
- the effects and justifiability of interests based on these scores.

Frederiksen and Collins (1989: 27) point out that a test (or examination) may be considered to be “systemically valid” when it “induces in the education system curricular and instructional changes that foster the development of the cognitive skills that the test is designed to measure”. This is also true of assignments in both summative and formative assessments. In essence, the concept relates to the notion of consequential validity and desirable washback. If, however, the test, examination or assignment has a negative effect on the development of the abilities it is purportedly designed to measure, the validity argument is weakened. Once again, this was a crucial aim of this thesis, particularly with reference to the varied characteristics of the target group and the particular exigencies of the ODL context.

As Du Plessis (2012: 25) observes, the validation process can be described as referring to the “systematic presentation of this evidence as a unity within a multiplicity of arguments” illustrating the relationship of the... examination to the definition of the construct being assessed”. The content and the assessment criteria should be based on the definition of the construct, which is of central importance to the validation process.

It is therefore vital that the features characterising the construct be clearly and specifically defined in order to ensure validity.

The construct measured by an assessment is found in the interaction between underlying ability, context of assessment and the scoring process (i.e. cognitive validity, context validity and scoring validity), a symbiotic relationship described by Shaw and Weir (2007) as the “constructed triangle”. This triangulation design strengthens the validity argument because it demonstrates the evidence in a number of ways, thus making it easier to determine whether the assessment is appropriate for the intended purpose (Sharton 1996: 68). This valuable concept was borne in mind in the present study where evidence of scoring validity amongst the markers was influenced by the formulation and clarity of the construct.

It should be cautioned that: “Evidence for different validities does not have to be equal in amount or strength, since an increase in one type of validity necessarily results in increased overall validity” (Hattingh 2009: 57). According to Hattingh, “the most secure support for a validity argument” is the collection of “various types of validity evidence, with a focus on the most relevant types”.

4.3.2 Challenges relating to determining the construct for ENG1501.

As noted in Chapter 2, determining the construct for the target module (ENG1501) presented some difficulty since certain criteria for the specific outcomes lack clarity, particularly in the case of the first assessment criterion for the second outcome (Appendix C). It can also be argued that this second specific outcome is repetitive (in addition to being poorly articulated). The first specific outcome (“Read a range of literary texts in different genres [poetry, prose and drama] with comprehension at an inferential level” and to apply “[a]ccepted conventions of academic discourse”) is extrapolated from the general statement given in the summary of the outcomes, namely that:

Students credited with this module will be able to apply appropriate reading strategies to a wide variety of literary and non-literary texts in English. They will also be able to demonstrate basic skills of writing academic English.

The second specific outcome (“Demonstrate basic awareness of the creative choices made by writers of literary texts in English”) is problematic, especially regarding the second criterion, (“The dimensions of artistry and contrivance in the composition of literary texts in English are explored and explained through acceptable academic discourse”). As stated in Chapter 2, the wording of this criterion is vague and lends itself to possible misinterpretation. This gives rise to the problem of identifying the construct and operationalising it for the purpose of validation. It could also be argued that the outcome statement is out of touch with a large number of mainly L2 students whose motivation is instrumental rather than integrative (Butler 2006: 113), and for whom “artistry” and “contrivance” are not necessarily goals towards which they strive.

In the case of the existing assessment scale (Appendix B), attempts were made to extrapolate guidelines for the description of an “excellent” answer. These criteria for “clearly demonstrating the skills required by the NQF [National Qualifications Framework] criteria” are summarised as follows:

- familiarity with – recognising and recalling – the subject matter;
- understanding [the subject matter];
- application of this information;
- analysis, for instance of relationships;
- evaluation (for example critiquing different approaches).

According to the rating scale, the organisation of the essay should be:

- focused on assigned topic;
- thoroughly developed.

It could be argued that the criteria for organisation form part of “acceptable academic discourse” as implied in the second specific outcome. Other requirements of academic writing, reflected in the rating scale, deal specifically with form. This scale concentrates on vocabulary, language usage and mechanics.

Despite the attempt to extrapolate criteria from the given outcomes, the concern remains that certain criteria stated in the present assessment scale remain generalised, vaguely worded and, thus, open to differing interpretations, especially in the distance learning context. It should be noted that the criteria are cited in the rating scales of other modules that assess academic writing but do not have a literary component. A further observation is that, in this generalised rating scale, the specific requirement of the construct is not emphasised sufficiently, namely, that students demonstrate knowledge and understanding of the prescribed literary texts and can “apply appropriate reading strategies to a wide variety of literary and non-literary texts in English”.¹⁰ The question that arises is: to what extent do the criteria given in the rating scale match the stated outcomes of the course.

A practical problem may also be encountered in assignment marking as the criteria are mentioned only in connection with the category: “excellent” and, although it is assumed that the criteria will be applied to other categories, the danger is that they can easily become lost in the process of assigning a mark at these levels. In line with a normal Bell curve, the students falling into the category “excellent” will represent a minority of those registered for the course. Furthermore, the balance between subject knowledge and academic writing style needs to be examined and interrogated more closely. For example, care should be taken not to over-emphasise language skills at the expense of the student’s knowledge of and insight into the given text. These hypothesised weaknesses have been addressed in the empirical section of this thesis, both in the consideration of the quantitative data, and qualitatively in the form of comments and questionnaires (Chapters 6 and 7).

Furthermore, the context of assessment for ENG1501, as described in Chapter 5, is challenging in terms of the heterogeneous composition of the target group and the geographical distance between stakeholders. Thus, assessment takes place in a multilingual and multicultural context, representing a wide age range and differing socio-economic circumstances (Section 5.6.1). These factors pose difficulties when researchers or lecturers attempt to design an assessment that is fair to all test-takers. An additional problem (as mentioned in Chapter 2) is that the module was not conceived

¹⁰ In the case of ENG1501, no “non-literary texts” are examined.

originally with an ESL target group in mind and is now attempting to cater for English Home Language speakers as well as those for whom English is a second (and sometimes third or fourth) language in a multilingual repertoire.

Du Plessis and Weideman (2014: 128) point out that this problem is exacerbated in the South African school environment by the fact that students registered for English Home Language are not necessarily HL speakers. Du Plessis and Weideman (2014: 128) add that it cannot be assumed, therefore, that spoken and written proficiency will be on a L1 level, and also that it “is becoming increasingly difficult” to distinguish between the various levels (i.e. English First, Additional, Second or Foreign language) in a multilingual environment such as South Africa. It should also be noted that it cannot be taken for granted that all ESL speakers possess equal levels of English proficiency because some could have attained nearly L1 fluency as a result of constant exposure to the language – for example, at school and in the workplace. As discussed in Chapter 2, these conditions present challenges to language teaching and research in the ODL tertiary context, and have to be considered in the validation process of the rating instrument of the target module, as well as during the design of an alternative scale.

Scoring validity should take into account scoring criteria and the rating scale and process. This includes assessor characteristics, the awarding of marks and post-examination feedback. Assessor characteristics, as well as the danger of inconsistency and subjectivity, have been discussed in the consideration of factors that have an impact on scores (Section 4.6). In the case of the target module, moderation took place in order to attain a balanced assessment and to mitigate the potential danger of varying interpretations of the scale. It is noted, however, that some marks are moderated electronically and not solely by the module co-ordinator or second examiner. The reason for this is that, given the large number of students, manual moderation by second examiners became simply too unwieldy and time-consuming. However, in ENG1501, this is done only after a percentage of the responses have been quality controlled by moderators. In some courses, moderation is carried out by setting certain parameters for the electronic moderation of marks. There is an automatic adjustment where 48 and 49 are moved to 50 (a pass mark) and where 73 and 74 are upgraded to 75 (a distinction). It is not yet apparent whether this system results in greater objectivity or whether it leads to an inflexible approach (in some cases) devoid of human influence. It is debatable

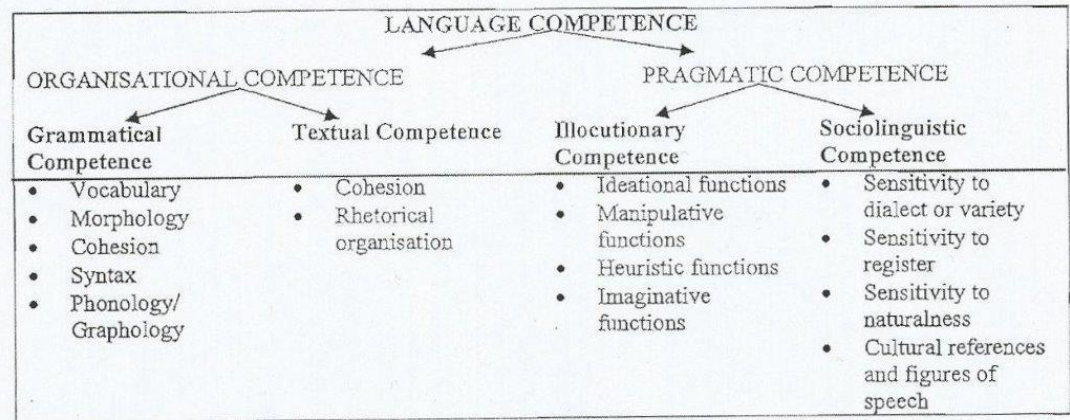
whether this automatic adjustment can be equated with a considered appraisal to determine whether the balanced assessment has taken place and the scale has been interpreted accurately. This is particularly relevant in the study of English literature, which encourages divergent answers and perspectives and rightfully gives rise to very different, but sometimes equally correct answers. It could be argued that the pre-marking training is more effective in achieving this.

4.4 MODELS AND FRAMEWORKS

In order to explain the distinction between the terms “model” and “framework”, Fulcher and Davidson (2007: 36) describe models as “overarching and relatively abstract theoretical descriptions of what it means to be able to communicate in a second language”. A framework is defined as “a selection of skills and abilities from a model that are relevant to a specific context” (Fulcher & Davidson 2007: 36). Thus, a model could be described as demonstrating the cognitive processes of writing, whereas a framework suggests the evidence to be collected as well as the manner in which it can be obtained.

4.4.1 Bachman’s model of communicative ability

The highly influential model of communicative ability designed by Bachman (1990), shown in Figure 4.2, and its later adaptation by Bachman and Palmer (1996) forms the foundation for validation frameworks such as the Cambridge VRIP framework, Weir’s socio-cognitive framework and the interactionist framework of Shaw and Weir (2007).



Source: (Bachman 1990: 87)

Figure 4.2: Components of language competence

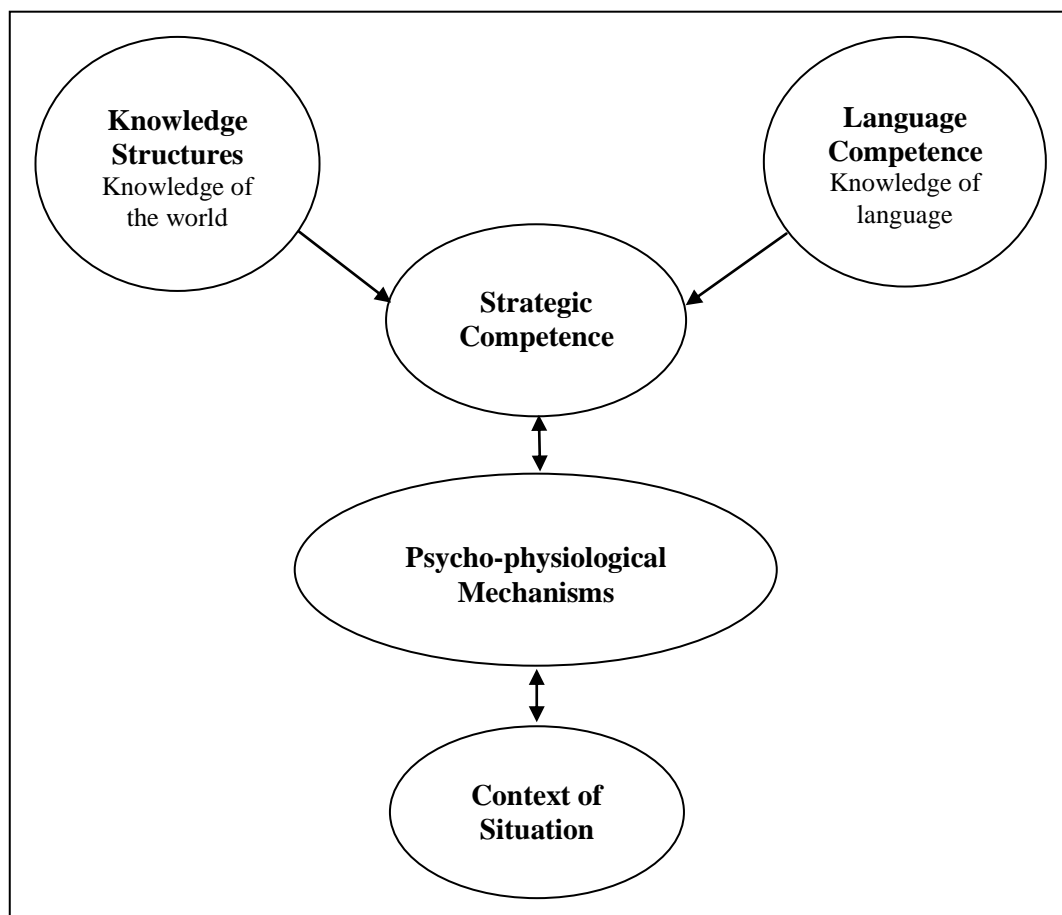
Bachman (1990; 2005) stresses that language assessment validation should be based on two stages, which he describes as the “what” and the “how” of language testing. The “what” refers to the abilities to be measured, and includes the attributes brought into the testing situation by the test-takers. Bachman labels the language abilities that are measured by language tests as language competence, strategic competence and psychological processes. This categorisation also takes into account the interaction between these components, as well as between abilities and the context in which these are tested (Bachman, 1990: 87). The scale proposed by the present study attempts to address this interaction by considering the varied language competences of first-year students in the very complex ODL context, and how these can be addressed by a fair assessment process.

The second stage (“how”) of the model comprises the facets of the test method (Bachman, 1990: 81). According to Bachman (1990: 81), test method features that might influence performance are the:

- testing environment;
- testing rubric;
- available input;
- nature of expected response;
- relationship between input and response.

A major component identified by Bachman (1990: 81) is communicative language ability, which can be observed only indirectly. This poses a challenge to the formulation of a construct because inferences must be made based on test performances (Bachman, 1990: 256). A difficulty arose in the case of the target group since initial records of performance were based on Grade 12 marks, which reflect a generalised language competence at the school-leaving level, and might not be appropriate to the level required by ENG1501, especially if the student was enrolled at school level for English as an “additional language” (i.e. in the case of second language students). As the tables provided in Section 5.6.1 demonstrate, the majority of students enrolled for the module are additional language (AL) speakers of English.

Bachman (1990: 85) maintains that the context of language use and the content of the rating scale should be examined in the light of the general model of language ability as depicted in the graphic representation below. Bachman (1990: 85) adds the component of strategic competence, which refers to the cognitive ability to relate language competence to the performance context – in other words, the ability to plan, assess and execute language performance to attain a communicative goal.



Source: (Bachman 1990:85)

Figure 4.3: Components of communicative language ability in communicative language use

Bachman's views have been praised for providing the "conceptual foundation" (Bachman 1990: 1) of language testing, and as "the most influential mark of the 1990s" (Chapelle, 1999: 257). McNamara (2003: 466) agrees that "the publication of Bachman (1990) was a major event" and considers Bachman's model of communicative language ability to be of particular significance. However, McNamara and Roever (2006: 32) take issue with Bachman's consideration of the social context of language use, which they

believe is inadequate. McNamara and Roever (2006: 32) believe that the characterisation of the social dimension in terms of individual ability “severely constrains the conceptualization of the social dimension of language testing context”, and caution that Bachman’s (1990) model is not “a theory of social context in its own right” (McNamara & Roever 2006: 32). Although this caution is valuable, it can be argued that Bachman’s theory provides a very useful foundation on which further research can be based. It is notable that Bachman (1990: 1) used the term “foundation” in describing his concept, and did not claim it to be “a theory of social context in its own right” (McNamara & Roever 2006: 32).

4.4.2 Bachman and Palmer

Bachman and Palmer’s model (1996) is based on that of Bachman (1990). The model’s most important adaptation of Bachman’s (1990) model is the replacement of construct validity with test usefulness as a central criterion. Qualities that contribute to test usefulness are reliability, construct validity, authenticity, interactiveness, impact and practicality. A balance should be established between the elements of test usefulness according to the specific context and assessment situation. Like the Bachman (1990) model, Bachman and Palmer (1996) identify the grammatical, textual, functional and socio-linguistic components of language competence.

Unfortunately, the relationship between the qualities of test usefulness, as well as the link between test use and construct validity, seems unclear. Both models have also been criticised for their lack of a clear explanation of the interaction between the components of language competence (i.e. grammatical, textual, functional and socio-linguistic) during language use, particularly with respect to the function of grammatical knowledge in this process. Purpura (2004: 55) goes as far as to label Bachman and Palmer’s (1996) description of grammatical knowledge as unhelpful as it does not adequately assess form and meaning. This could present problems in the case of a target group such as ENG1501, where an over-emphasis on form at the expense of meaning could result in inaccurate scoring. Furthermore, as Hattingh (2009: 71) notes, “Although Bachman and Palmer’s (1996) notion of test usefulness suggests a different way of conceptualising validity and validation, it downgrades construct validity to an aspect of usefulness”.

Despite these reservations, the models of Bachman (1990) and Bachman and Palmer (1996) have been highly influential, laying the foundations for later frameworks such as the Cambridge ESOL (VRIP), Weir's framework (2005) and the framework of Shaw and Weir (2007), as well as adaptations by Bachman and Palmer (2010).

4.4.3 Cambridge ESOL framework

The Cambridge ESOL framework is based on what is described as the VRIP system. The acronym represents four criteria of test usefulness, namely validity, reliability, impact and practicality. Hawkey (2006: 18) points out that these criteria overlap to a large extent with the qualities proposed by Bachman and Palmer (1996) (i.e. reliability, construct validity, authenticity, interactiveness, impact and practicality).

In the Cambridge framework a balance is sought between VRIP and, thus, according to Saville (2003) and Hawkey (2006), a rating scale is envisaged that should:

- be appropriate for the purpose of the assessment;
- produce similar scores over repeated assessments;
- have a positive influence on individual stakeholders and the general education process;
- be practical to develop, produce and administer.

The ESOL framework is based on the unitary view of validity as proposed by Messick (1989). Messick highlights the equation of validity with fitness of purpose. Similarly, validity has been described as the central component of the VRIP qualities (Shaw & Jordan 2002: 11). The dominant principle of validity is seen as “fitness for purpose” and, thus, the emphasis is on content validity.

The ESOL framework has been criticised for over-emphasising practicality. Hattingh (2009: 76) states:

The VRIP framework also places much emphasis on practicality, although ease of use does not inherently affect validity. A valid instrument that cannot be implemented due to practical limitations, or needs particular resources (e.g. staff, time or finances) in order to be implemented, may not be of immediate use, but it remains a valid instrument.

A further criticism is that, while all of the VRIP elements are considered important in ascertaining the usefulness of an assessment instrument, achieving a balance between these elements may present difficulties. A revision of the framework was proposed in an article by Weir and Shaw (2005), which employed Weir's (2005) socio-cognitive framework to improve the VRIP process. This significant framework will now be discussed.

4.4.4 Weir's framework (2005)

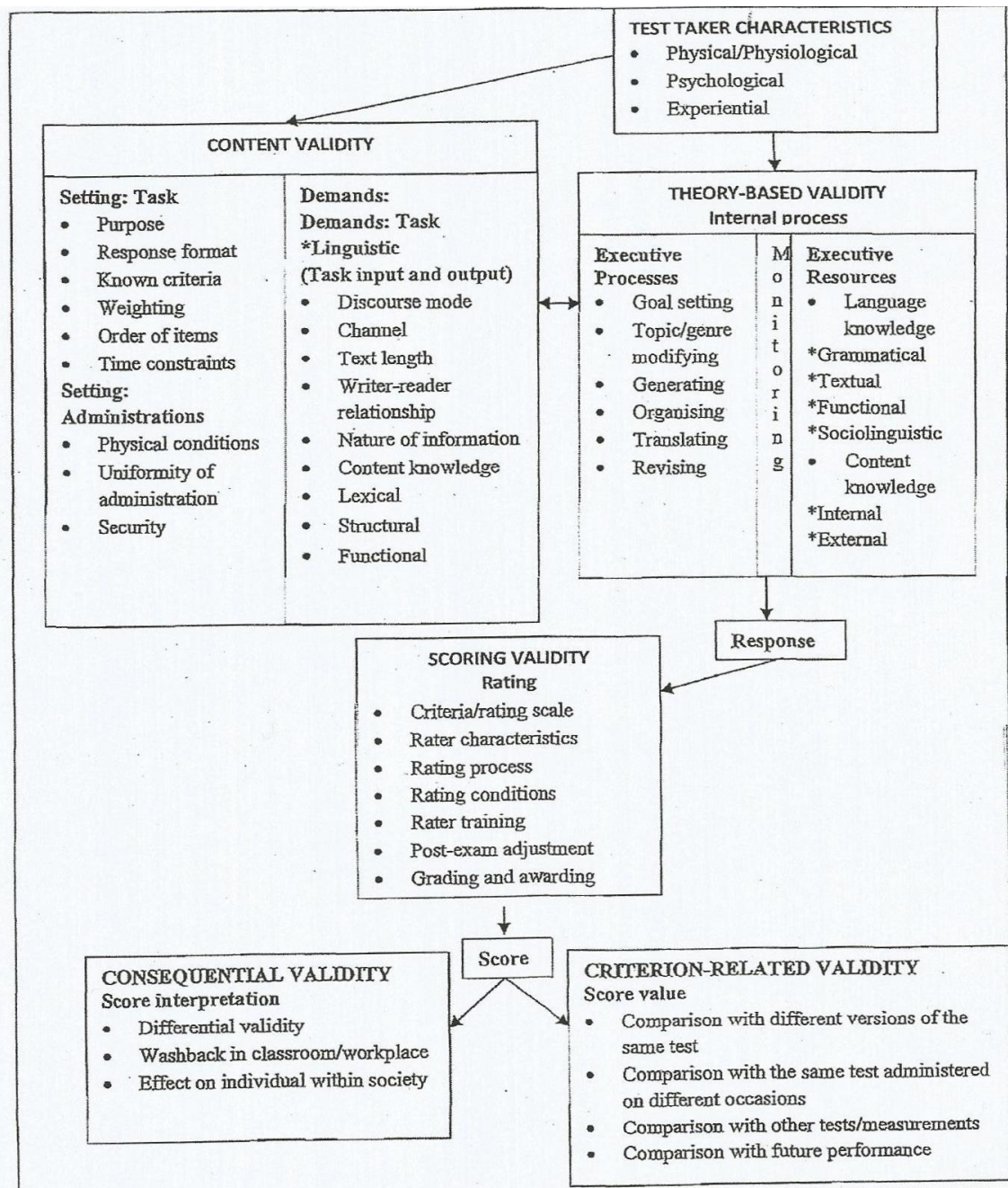
The significant frameworks of Weir (2005) and Shaw and Weir (2007) have been mentioned with reference to validity in Chapter 3. Weir (2005) emphasises the social aspects of language use, and advocates a socio-cognitive assessment framework that incorporates social aspects (such as context and audience) in addition to cognitive elements of assessment (Weir 2005: 12). The main elements of Weir's (2005) framework comprise context validity, theory-based validity, scoring validity, consequential validity and criterion-related validity. These types of validity are presented within a unified model, with construct validity as a super-ordinate category uniting the various elements.

The evidence generated by each of these elements supports the interpretation of test scores. Weir (2005: 47) explains that:

The more comprehensive the approach to validation, the more evidence collected on each of the components of this framework, the more secure we can be in our claims for the validity of a test. The higher the stakes of the test, the stricter the demands we might make in respect of all of these.

Weir (2005: 12) believes that the process of ascertaining validity requires "multifaceted and different types of scores on a test", and thus highlights scoring validity as a primary criterion for accurate assessment. With this in mind, Weir (2005: 49) cautions against over-emphasis on practicality at the expense of the construct, arguing that: "We should not consider method before trait".

A later article by Weir and Shaw (2005: 10) employed the socio-cognitive framework of Weir (2005), shown in Figure 4.4, to improve and renew the Cambridge VRIP process. This framework demonstrates Weir's belief in the equal importance of cognitive and social factors, as well as his related statement that test validity requires a multi-faceted approach to test scores (Weir 2005: 12). Weir's framework is helpful because it reflects a complex assessment process in which contextual and cognitive elements are identified and the relationship between them is taken into consideration (Weir, 2005: 43; Weir & Shaw, 2005: 11).

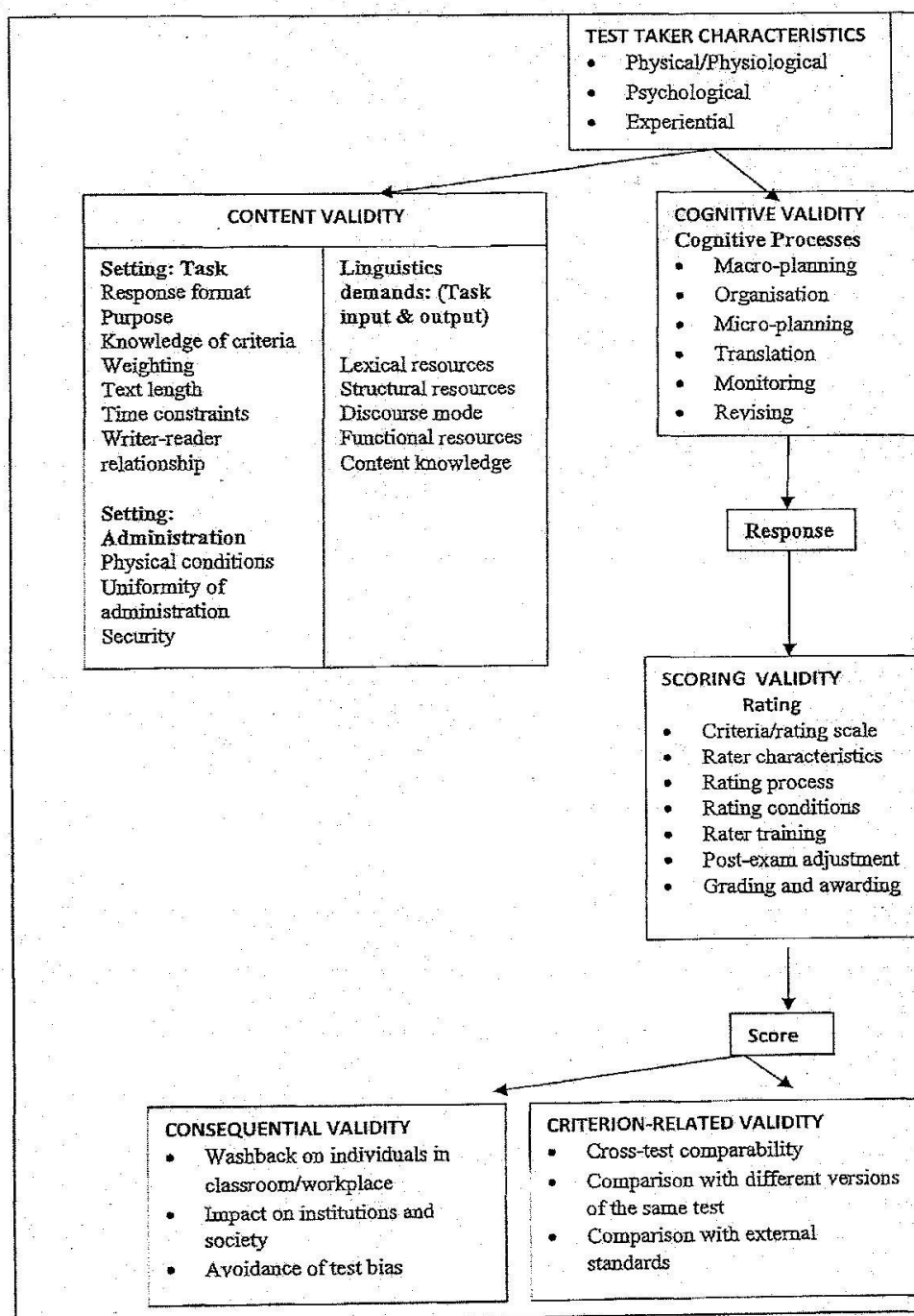


Source: (Weir 2005)

Figure 4.4: Weir's socio-cognitive framework

4.4.5 The framework of Shaw and Weir (2007)

The framework designed by Shaw and Weir (2007) is shown in Figure 4.5.



Source: Shaw and Weir (2007)

Figure 4.5: Validation framework designed by Shaw and Weir

4.4.5.1 Comparison of the two scales

Like Weir (2005), Shaw and Weir (2007) emphasise the importance of scoring validity (which encompasses all factors that have a direct influence on scores) and of a valid rating scale, which they consider essential to ensuring a valid and reliable assessment that reflects the particular construct being assessed. In agreement with Weir (2005),

Shaw and Weir (2007) regard construct validity as central to the relationship between other types of validity (context, cognitive and scoring validity). The framework presents two dimensions of a construct, namely the underlying cognitive ability and the context of use (task and setting). In addition, Shaw and Weir (2007) introduce the scoring dimension as a third important element, since scoring criteria describe the required performance level and, in conjunction with context, are thus essential to the formation of the construct.

A difference from Weir's (2005) framework is the replacement of theory-based validity with cognitive validity that refers to the cognitive processes that students engage in in order to respond to writing tasks. The multi-faceted concept of reliability is subsumed under the function of scoring validity and thus unites reliability with the concept of validity, instead of the traditional viewpoint in which reliability is considered to be a separate element of assessment, often in conflict with validity (Messick, 1989). These adjustments streamline the framework without altering Weir's (2005) original concept significantly.

Shaw and Weir (2007: 239) caution against an over-simplification of the framework. They point out the need for evidence that is relevant to the particular circumstances, particularly to the level of language ability that is being tested. However, the framework of Shaw and Weir is comprehensive, and provides a valuable guide in the validation process adopted by this study which deals with a complex, multi-faceted context. This will be interrogated further in the report on the empirical process of the current study. (Chapters 6-8).

4.5 CRITERIA AND BAND SCALES

Since the validity of an assessment scale depends on the degree to which it represents the construct being assessed, scoring validity must be ensured. To achieve this, assessment designers should reach consensus on the criteria to be tested and their implications.

It is thus essential that criteria should be clear and explicit (Weir 1990; 2005; Hamp-Lyons 1990; Bachman & Palmer 1996; Brown *et al.* 2004; Elder 2005; Hattingh 2009).

4.5.1 Level descriptors and band levels

In order to demonstrate scoring validity, criteria should be clearly articulated, explicitly described, and relevant to the construct. Furthermore, developers should reach consensus on the criteria to be assessed, as well as what these signify.

There is no ideal number of level descriptors to be included in a scale. These would depend on the construct to be assessed. A guideline would be that the scale should consist of sufficient criteria to reflect the construct and the student's abilities while, at the same time, remaining easy to use. However, one should consider the psychological effect that too many criteria could have on stakeholders. This is of particular importance in the ODL context, where markers are dealing with large numbers of students, are working under time pressure, and where there is no classroom post-assessment that would assist in clarifying misinterpretations or identifying gaps in the rating scale or in the students' grasp of the subject matter.

With this guideline in mind, researchers have suggested a limit of seven criteria (Council of Europe 2001; Luoma 2004; Weir 2005), with Luoma (2004:80) preferring a maximum of five to six criteria. In order to ensure that criteria are comprehensive, Weir (1990: 68) believes that they should be based on empirical evidence obtained from sample scripts. This advice was followed in the current research.

Chapelle (2012: 22) agrees with this procedure but cautions that "the precise methods for sampling in both corpus compilation and test task development need to be better specified and evaluated if they are to be used as support for a sampling inference in the validity argument". Thus, it is important that descriptors are expressed precisely, and that the distinction between them is clear to enable assessors to assign scores accurately (Bachman 1990: 36). For this reason, it is also advisable to avoid describing levels by comparing them to one another (for example, "better than..." or "poorer than..."). Descriptors of this kind could lead to confusion and thus inconsistency on the part of the

assessors (Weigle, 2002: 125). As Hattingh (2009: 145) observes, “descriptions must be unambiguous and give raters a specific indication of how criteria are manifested (in terms of salient features) at each performance level”. For example, the Common European Framework of Reference for Language (CEF: 2001), developed by the Council of Europe, shown in Table 4.1 below, provides the following descriptors for two of its six scales (Hudson, 2005: 217).

**Table 4.1: Common European framework (Table vi) – global scale
(Council of Europe, 2001)**

Proficient User	C1	Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic, and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors, and cohesive devices.
...
Basic User	A2	Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.

Source: Adapted from Hattingh (2009:145)

An excerpt from the International English Language Testing Systems (IELTS) scale, shown in Table 4.2 below, also provides a clear explanation of criteria, in this case for Bands 9 and 8, in descending order of proficiency.

Table 4.2: IELTS writing band level descriptors for bands 8 and 9

Band	Task achievement	Coherence and cohesion	Lexical resource	Grammatical range and accuracy
9	<ul style="list-style-type: none"> fully satisfies all the requirements of the task clearly presents a fully developed response 	<ul style="list-style-type: none"> uses cohesion in such a way that it attracts no attention skillfully manages paragraphing 	<ul style="list-style-type: none"> uses a wide range of vocabulary with very natural and sophisticated control of lexical features; rare minor errors occur only as 'slips' 	<ul style="list-style-type: none"> uses a wide range of structures with full flexibility and accuracy; rare minor errors occur only as 'slips'
8	<ul style="list-style-type: none"> covers all requirements of the task sufficiently presents, highlights and illustrates key features/bullet points clearly and appropriately 	<ul style="list-style-type: none"> sequences information and ideas logically manages all aspects of cohesion well uses paragraphing sufficiently and appropriately 	<ul style="list-style-type: none"> uses a wide range of vocabulary fluently and flexibly to convey precise meanings skillfully uses uncommon lexical items but there may be occasional inaccuracies in word choice and collocation produces rare errors in spelling and/or word formation 	<ul style="list-style-type: none"> uses a wide range of structures the majority of sentences are error-free makes only very occasional errors or inappropriacy

Source: IELTS Table

Table 4.3 shows a summary of the number and nature of the criteria for four major rating scales.

Table 4.3: Summary of criteria distinguished in four current rating scales

Scale	Number of criteria	Nature of criteria
Jacobs et al. (1981)	5	<ul style="list-style-type: none">• content• organisation• vocabulary• language use• mechanics
Cambridge Main Suite Examination	4	<ul style="list-style-type: none">• fulfilment of the task set• communicative command of the target language• organisation of discourse• linguistic errors
International English Language Testing System (IELTS) Writing Assessment scale	4	<ul style="list-style-type: none">• task achievement/response• coherence and cohesion• lexical resource• grammatical range and accuracy
TEEP Attribute Writing Scales (Weir, 1983)	7	<ul style="list-style-type: none">• accuracy• fluency• interaction• coherence and organisation• task fulfilment• language control and linguistic range• communicative effectiveness• register

Source: (Hawkey & Barker 2004: 123)

4.5.2 Band levels

The choice of the number of band levels depends on largely practical considerations (McNamara 2000: 42; Luoma 2004: 80). McNamara (2000: 42) points out that there is “no point in proliferating descriptions outside the range of ability of interest. Having too few distinctions within the range of such ability is also frustrating, and the revision of rating scales often involves the creation of more distinctions”. These considerations might have led to the creation of scales such as the IELTS Tables iv and v (See Table 4.4 below) that show distinctions (or half-levels) within a particular level. These allow for finer and more precise distinctions within levels. In effect, the IELTS scale distinguishes nine levels in contrast to the scale of Jacobs *et al.* (1981), on which the ENG1501 scale was based, which distinguishes only four levels.

Table 4.4: IELTS band scale descriptors

IELTS Band Scale
Band 9 – Expert User Has fully operational command of the language: appropriate, accurate and fluent with complete understanding.
Band 8 – Very Good User Has fully operational command of the language with only occasional unsystematic inaccuracies and inappropriacies. Misunderstandings may occur in unfamiliar situations. Handles complex detailed argumentation well.
Band 7 – Good User Has operational command of the language though with occasional inaccuracies, inappropriacies and misunderstandings in some situations. Generally handles complex language well and understands detailed reasoning.
Band 6 – Competent User Has generally effective command of the language despite some inaccuracies, inappropriacies and misunderstandings. Can use fairly complex language, particularly in familiar situations.
Band 5 – Modest User Has partial command of the language, coping with overall meaning in most situations, though is likely to make many mistakes. Should be able to handle basic communication in own field.
Band 4 – Limited User Basic competence is limited to familiar situations. Have frequent problems in understanding and expression. Is not able to use complex language.
Band 3 – Extremely Limited User Conveys and understands only general meaning in very familiar situations. Frequent breakdowns in communication occur.
Band 2 – Intermittent User No real communication is possible except for the most basic information using isolated words or short formulae in familiar situations and to meet immediate needs. Has great difficulty in understanding spoken and written English.
Band 1 – Non user Essentially has no ability to use the language beyond possibly a few isolated words.

Source: (IELTS, 2007:4)

By comparison, the scoring profile of Jacobs et al. (1981) distinguishes five criteria with varying weights and four band levels, as shown in Table 4.5 below:

Table 4.5: The Jacobs' scoring profile

ESL COMPOSITION PROFILE			
STUDENT		DATE	TOPIC
SCORE	LEVEL	CRITERIA	
CONTENT	30 - 27	EXCELLENT TO VERY GOOD: knowledge; substantive; thorough development of thesis; relevant to assigned topic	
	26-22	GOOD TO AVERAGE: some knowledge of subject; adequate range; limited development of thesis; mostly relevant to topic, but lacks detail.	
	21-17	FAIR TO POOR: limited knowledge of subject; little substance; inadequate development of topic.	
	16-13	VERY POOR: does not show knowledge of subject; non-substantive; not pertinent; OR not enough to evaluate	
ORGANISATION	20-18	EXCELLENT TO VERY GOOD: fluent expression; ideas clearly stated/supported; succinct; well-organised; logical sequencing; cohesive.	
	17-14	GOOD TO AVERAGE: somewhat choppy; loosely organised but main idea stands out; limited support; logical but incomplete sequencing.	
	13-10	FAIR TO POOR: non-fluent; ideas confused or disconnected; lacks logical sequencing and development.	
	9-7	VERY POOR: does not communicate; no organisation; OR not enough to evaluate.	
VOCABULARY	20-28	EXCELLENT TO VERY GOOD: sophisticated range; effective word/idiom choice and usage; word form mastery; appropriate register.	
	17-14	GOOD TO AVERAGE: adequate range; occasional errors of word/idiom form, choice, usage but meaning not obscured.	
	13-10	FAIR TO POOR: limited range; frequent errors of words/idiom form, choice, usage; meaning confused or obscured.	
	9-7	VERY POOR: essentially translation; little knowledge of English vocabulary, idioms, word form; OR not enough to evaluate.	
LANGUAGE USE	25-22	EXCELLENT TO VERY GOOD: effective complex construction; few errors of agreement, tense, number, word order/function, articles, pronouns, prepositions.	
	21-18	GOOD TO AVERAGE: effective but simple construction; minor problems in complex construction; several errors of agreement, tense, number, word order/ function, articles, pronouns, prepositions but meaning seldom obscured.	
	17-11	FAIR TO POOR: major problems in simple/complex constructions; frequent errors of negation, agreement, tense, number, word order/ function, articles, pronouns, prepositions and/or fragments, run-ons, deletions; meaning confused or obscured.	
	10-5	VERY POOR: virtually no mastery of sentence construction rules; dominated by errors; does not communicate; OR not enough to evaluate.	
MECHANICS	5	EXCELLENT TO VERY GOOD: demonstrates mastery of conventions; few errors of spelling, punctuation, capitalisation, paragraphing.	
	4	GOOD TO AVERAGE: occasional errors of spelling, punctuation, capitalisation, paragraphing.	
	3	FAIR TO POOR: frequent errors of spelling, punctuation, capitalisation, paragraphing; poor handwriting; meaning confused or obscured	
	2	VERY POOR: no mastery of conventions; dominated by errors of spelling, punctuation, capitalisation, paragraphing, handwriting legible; OR not enough to evaluate.	

Source: (Jacobs et al. 1981)

The scale shown in Table 4.5 was modified and adapted into the scale used for ENG1501 and other modules at Unisa. The current scale distinguishes between organisation/content and language use (see Appendix B). This has simplified the marking process, but the question can be asked if over-simplification has not occurred in some instances, particularly in the case of Level 3 (Language Use) as shown in Table 4.6:

Table 4.6: Language Use Level 3 of the scale used for the target module

13-8 (54%- 32%)	3 FAIR TO SHAKY: AT RISK	Vocabulary: small range, frequent issues of word/idiom, choice, usage Language usage: major problems in simple/complex constructions, frequent language issues including sentence construction problems, meaning confused or obscured Mechanics: frequent problems with mechanics, untidy handwriting, meaning confused or obscured
-----------------------	---	--

Source: Adapted from (Jacobs et al. 1981)

The problem with Level 3 is that the range is too wide, especially considering that the pass mark is 50%, a mark that occurs in the middle of the range but with no indication of what should distinguish between a ‘pass’ mark and a ‘failure’. It is a matter of concern that subjective or inconsistent marking could jeopardise the fairness and reliability of the scale, and could potentially result in serious consequences in the case of students who fail because of being placed under the 50% mark. The question arises as to why a script deemed to have achieved a pass mark i should share the assessment level and assessment criteria with those labelled “at risk”.

Addressing the issue of band levels, Council of Europe (2001: 21) advises that the “number of levels adopted should be adequate to show progression in different sectors, but, in any particular context, should not exceed the number of levels between which people are capable of making reasonably consistent distinctions”. As demonstrated by the scales discussed in this section (4.5.2), writing scales may differ in the number of levels, and still meet the criterion posited by the Council of Europe (2001: 21).

4.6 FACTORS THAT IMPACT UPON SCORES

Huang (2009) categorises factors that impact on scores into two broad types: rater-related and task-related. The rater-related category includes the ranking method used, rating criteria, raters’ academic disciplines, professional experiences, linguistic backgrounds, tolerance for error, perceptions and expectations, and rater training. Task-related factors include the types of writing tasks and their difficulty levels.

In the current thesis, a broader categorisation was adopted. The first group includes factors of primary importance to effective assessment and directly related to learning,

teaching and assessment. These include the characteristics of the test-takers, inter-rater variance, and the washback effect. A secondary group includes the impact of administrative and physical factors such as the administrative setting, uniformity of administration, security, and the physical environment of testing.

4.6.1 Factors directly related to learning, teaching and assessment

These factors include those that are related to the pedagogy of language teaching and learning, such as test-taker characteristics, inter-rater variance, and the washback effect.

4.6.1.1 *Test- taker characteristics*

Test-takers are indubitably the main component of any assessment process and, thus, factors that affect their performance should be borne in mind when designing the content of an assessment scale. Hattingh (2009: 92) notes that: “Test-takers use resources such as their content knowledge, which may be existing background knowledge or provided by the task input... when responding to a task. These factors determine how learners approach, plan and execute a task”. Various factors such as age, interest, experience, knowledge and motivation affect test-takers' performances.

O’Sullivan (2006: 3) maintains that a scale that has been developed against the background of test-takers’ characteristics is less likely to demonstrate bias. Unfortunately, as O’Sullivan notes also, test designers tend to develop assessments according to their own perceptions of the target population instead of empirical evidence about the group.

According to O’Sullivan (2000; 2006), test taker characteristics can be divided into three categories. These are physical/physiological, psychological and experiential. Physical/physiological characteristics include short-term illnesses, disabilities, age and gender. Psychological characteristics encompass memory, cognitive style, motivation, concentration and emotional state. Experiential characteristics refer to factors relating to the test-takers’ previous experience, particularly their former education, examination experience and communicating in an L1 environment. These three categories of test

taker characteristics were later adopted by Shaw and Weir (2007) and incorporated in their framework.

Bachman (1990: 146) makes a distinction between systematic and unsystematic characteristics. The former include students' content knowledge, cognitive style, age, gender and physical disabilities. These factors consistently influence students' performances in the same way and can be controlled to a degree. On the other hand, unsystematic characteristics are random, unpredictable and often temporary influences (such as personal circumstances and emotional issues) beyond the control of test developers.

Ellis (1994) distinguishes between social context and social factors. Social context includes the settings in which formal and informal learning takes place, whereas social factors refer to age, gender, social class and ethnic identity. A combination of various social factors with different social contexts can influence results in a number of ways (Ellis 1994: 197).

The large number of variables identified as potentially influencing test-takers exacerbates the difficulty of assessment development, particularly in the case of the heterogeneous ENG1501 target group, which comprises students from various cultural, ethnic and language backgrounds, a wide age range and differing secondary educational environments (Chapter 5 Section 5.6.1. to 5.6.3). In the absence of geographical proximity of the stakeholders and the concomitant lack of face-to-face contact, these individual factors are not known to the lecturers and assessors of the module. This ODL context makes it difficult to test for systematic characteristics of the target group in order to ascertain the specific social contexts and factors involved. However, the demographic information presented in Section 5.6.1 provided information on the wide geographical, linguistic and ethnic ranges involved.

Shaw and Weir (2007: 19) advise that, given the exigencies of the subject content, it is extremely difficult to cater for all individual and specific needs and also meet the requirements of fair assessment. This is because of the wide range of variables that could influence test-takers. As noted, the problem is aggravated in large-scale assessments comprising heterogeneous groups, such as that of ENG1501. Every effort

should be made to prevent any student from being disadvantaged by the socio-cultural content of the assessment (Shaw & Weir 2007: 19). However, this is difficult in the case of the ENG1501 module, which deals with English literary texts and thus cannot avoid a socio-cultural content. In this respect, in ENG1501, an attempt was made to reflect the socio-cultural values of the majority of the students while, at the same time, introducing them to literary texts from a wider international canon (as seen in the prescribed texts listed in Appendix C). Furthermore, the recommendations made by Shaw and Weir (2007: 19) that tasks should reflect real-life communication and neutral topics would be easier to implement in modules dealing with generalised language and communication skills than in ENG1501. In this context, Weir (2005: 54) offers more practical advice by suggesting that every attempt should be made to ensure that test-takers are familiar with the types and features of assessment. This should include formative assessment or exercises with the same format as the final examination. These could be addressed by the online tutors and by podcasts presented by lecturers. Specimen papers could also be provided so that the students can work through these prior to the final assessment. These interventions would familiarise the students with the assessment criteria and content. Some of these interventions do occur in ENG1501, particularly in the form of study materials and the provision of past examination papers (although the latter is in the balance given copyright issues). Furthermore, students are provided with the assessment grid in their first tutorial letter.

4.6.1.2 *Inter-rater variance*

Not only do raters and test-takers frequently not share the same background, but it is often the case that inter-rater variance is caused by differing cultural influences. This is exacerbated in ODL, because raters are sourced throughout the country and seldom come into contact with one another. This prevents discussion and the sharing of perspectives that could lead to consensus on common assessment practices. However, training takes place at the beginning of each course, as well as at the start of the examination marking process, and this gives markers the opportunity to communicate and reach consensus on the examples of scripts provided. This minimises the problem, but the relative lack of communication caused by distance remains a problematic issue.

Furthermore, Wolfe *et al.* (1998) discovered that although assessors might appear to share a general understanding of scale content, they often seem to apply this scale content in different ways. For instance, more experienced and proficient raters tend to first read and then evaluate the student's work, whereas less proficient raters adopted an iterative approach, reading, re-reading and constantly referring to the scale. In addition, raters might have differing expectations and divergent socio-cultural backgrounds, and these could influence their assessments (Shaw & Weir 2007: 172).

According to Hamp-Lyons (1991: 242), raters should be aware of the elements of written discourse, as well as the multi-dimensionality of ESL writing in order to achieve a reasonably balanced score. This entails understanding the rating scale and specifying how scores should be assigned. Once again, consensus on these issues is more easily reached in FTFL than in ODL with its limited contact between raters.

Rater training helps to minimise the problem but, as Hattingh (2009: 153-154) points out, "the scoring process is still fairly unexplored, and too little is known about what goes on in raters' minds while scoring, about the effects of rater training, and the value of standardization". Researchers such as Huot (1990: 258), Lumley (2002: 246) and O'Sullivan (2006: 186), also comment on the dearth of knowledge about how raters reach their decisions during the scoring process. In this regard, Brindley (1989: 65) points out that one should not assume that raters' interpretation of a scale will be similar, even if that scale is valid, the criteria are clearly articulated, and the assessment process is well-organised. For example, McNamara (1996) discovered that, sometimes, raters think that they are scoring according to the given criteria, whereas in fact they are unintentionally not doing so. McNamara (1996) employed the multi-faceted Rasch scale to investigate rater behaviour in scoring writing assignments that were evaluating communicative skills. He found that, despite the fact that grammar was not emphasised, raters were unwittingly excessively influenced by grammatical accuracy. A similar finding was recorded by Azizi and Majdeddin (2014: 337) in a study attempting "to examine whether raters are actually assessing test-takers' writing samples based on the constructs defined in the scoring rubric" for IELTS writing skills. In the present study, this question was included in the questionnaire issued to markers.

However, Lumley (2002: 246) believes that "despite this tension and indeterminacy, rating can succeed in yielding consistent scores provided raters are supported by adequate training, with additional guidelines to assist them in dealing with problems". In other words, the aim of training is to ensure that raters agree with one another and apply the criteria consistently and as objectively as possible (Wolfe *et al.* 1998: 485).

4.6.1.3 Washback

Washback can be described as the effect of tests on teaching practice or, more specifically the influence of assessment on teaching, teachers and test-takers. The impact of assessment can also be widened to include the community at large (Shaw & Weir 2007).

Cheng (2008: 26) explains washback as the notion that the "test should drive teaching and hence learning". This might lead to teaching and learning being focused on areas that are likely to appear in tests or examinations. Washback might be intended or unintended and have positive or negative effects on the participants. Greene (2007) is of the opinion that, if an assessment develops the learner's overall abilities, washback is beneficial but, if this is not the case, washback could have negative consequences. Similarly, Hattingh (2009: 19) cautions against an assessment instrument that either "over- or under-represents a certain aspect of the construct domain", pointing out that this "assessment may lead to invalid scores, unfair inferences and negative washback effects". On the other hand, Hattingh (2009: 125) points out that if the test and assessment scale reflect the construct adequately, educators "can use the information from assessments to structure their practices and learners get an indication of which areas they still need to work on before mastering the skill. Detailed rating scales are useful for this purpose".

However, determining the effect of washback is not as simple as merely deciding whether its influence on teaching practice is positive or negative, depending on the content of the assessment or scale. Aldersen (2004: ix) describes the issue as "hugely complex". This is especially relevant in relation to student motivation (Fulcher & Davidson 2007: 222 - 224). The problem is that, although the extrinsic motivation envisaged by the assessment might encourage some students to achieve better results, it

might also cause anxiety, resulting in students performing below their level of intrinsic ability. Another problem is the danger of the educator teaching to the test (and thus narrowing the scope of the instruction) in order to obtain good results (Fulcher & Davidson 2007: 222 - 224). This would apply particularly to situations in which the stakes are high, and students are instrumentally motivated.

Hughes (2003: 53-57) provides the following guidelines to achieve positive washback:

- The abilities being tested should be those you want to encourage.
- Sample widely and unpredictably.
- Use direct testing.
- Use criterion-referenced assessment.
- Make sure test-takers and teachers are familiar with and understand the content.
- Assist educators where necessary by means of training and support materials.

Generally, washback should have a positive effect if educators use formative assessment results to identify and address students' needs. This highlights the importance of feedback that will assist students to improve their performance. As has been stated (Chapter 2), feedback is a challenge in ODL, but this challenge can be mitigated by using effective feedback strategies that should be designed for this environment. Feedback is seen as supporting learning and the challenges of providing this formative feedback are "multiplied in an ODL context because personal contact between students and lecturers is limited or non-existent" (Lephalala & Pienaar 2008: 69).

4.6.2 Administrative setting

The setting and other circumstances under which the assessment is administered can influence its validity (Weir 2005: 82; Shaw & Weir 2007: 133). Factors that can influence scores include physical conditions and procedures, uniform administration and security.

In the Unisa context, administrative aspects are challenged by the large number of examination venues, covering an extremely wide international network. Examination

centres are located throughout South Africa with 261 venues in various towns and cities, including 21 centres catering for prisoners. There are 59 venues in other African countries and 110 centres in countries outside Africa. This adds to the complexity of the distance learning environment, although every effort is made to ensure uniformity and to address any problems which might arise.

4.6.2.1 *Physical conditions*

Physical conditions include temperature, background noise and lighting. Ideally, these should be controlled to make sure that test-takers are as comfortable as possible. Unfavourable physical circumstances might have a negative effect on test-takers, causing them to under-perform, jeopardising the fairness of the assessment.

Physical conditions can be controlled in the case of summative assessment at Unisa, especially in the larger centres, where suitable lighting can be provided and temperatures can be adjusted as necessary through air-conditioning. It is possible that optimal conditions are not always present in smaller centres or in the case of certain international venues. However, investigations are carried out and problems addressed and prevented as far as possible by the administrative staff. Personnel at the venues are issued with instructions regarding physical conditions and other administrative issues pertaining to examinations.

In the case of the formative assessment assignments for ENG1501, it is not possible to control these factors because the student chooses (or is compelled by circumstances to use) a particular setting in which to write an assignment. These settings could vary greatly in terms of physical conditions and the physical safety of the students. However, final examination conditions can be controlled by the university because students write in examination venues under the surveillance of invigilators appointed by Unisa. Thus, any conditions or incidents affecting the physical well-being and safety of the candidates can be reported and addressed. If necessary, the examination can be rescheduled to take place under more favourable conditions.

4.6.2.2 Uniformity of administration

Test invigilators and other administrators should have clear and specific instructions and should adhere to these strictly. Uniformity is of the utmost importance because, if different test venues apply rules inconsistently (for example, by allowing more time than instructed), cognitive validity might be threatened because students' cognitive decisions might be influenced by such changes. At Unisa, factors are strictly controlled despite the problem of distance. Guidelines are given to invigilators of final examinations, although the large number and geographical spread of the venues (as described in 4.6.2) present a greater challenge than that encountered at FTF universities.

There is no supervision of the formative assessments as these are written in a venue of the students' choice, or as dictated by the students' circumstances. This is also the case in summative portfolio assessments (such as that for ENG1513). These contexts give rise to the danger of plagiarism and assistance by others and, thus, affect uniformity and cognitive validity negatively. In the case of ENG1501, students write a supervised examination at an examination centre, but plagiarism and assistance by others remain problematic in formative assessment, despite the fact that students are required to sign a form declaring that the assignment is their own work. Unfortunately, the university policy on plagiarism is not sufficiently specific, although actions that constitute plagiarism are explained and the students are warned about possible disciplinary action that might be taken in cases of plagiarism. Individual departments and even modules formulate their own assessment criteria and penalties for plagiarism, which may become increasingly severe depending on the student's level or the frequency of his or her plagiarism.

Cheng and DeLuca (2011: 117) found a "high co-occurrence of test structure and content and test administration/conditions". This finding has direct implications for test designers and administrators, as the testing administration/conditions greatly influence test performance. Cheng and DeLuca (2011: 117) add that "conditions under which a test is administrated could potentially alter the variance in the scores thus affecting validity".

In the case of the target module, the examinations and assignments are set by a panel of examiners. Marking is carried out by a team of markers (numbering from 15 to 30, depending on student numbers) and is moderated as described in Chapter 2. The same assessment scale is used for formative and summative assessments.

4.6.2.3 Security

Although access to the content of the formative assessments is available on the Unisa website and in the study material sent to students, the examination papers are never published or made accessible to stakeholders until the examination is written. If this security is breached, the results are irreparably skewed and validity is entirely jeopardised. In such cases, a new examination has to be administered.

The same potential problems apply to the security of the examination as were mentioned in 4.6.2.1 and 4.6.2.2. However, every effort is made to prevent a breach of security. Examination papers are sealed and transported under strict security measures, and invigilators are trained to familiarise them with the security procedures at all stages of the examination. Any security breach is reported to the University administration governing examinations. In the case of such a breach, examinations would have to be re-written. The University also works with law enforcement agencies in cases of examination leaks.

4.6.3 Impact on society

It is extremely difficult (if not impossible) to measure the impact of assessments on society in general (Weir 2005: 214). Despite this, Hamp-Lyons (1995: 299) believes that test developers should take into account the possible consequences of assessment on the broader society as well as the effect of washback on individuals. Hamp-Lyons (1995: 299) argues that assessment instruments should be evaluated from the perspective of all stakeholders, such as students, educators, parents, official bodies and potential employers. It should also be borne in mind that tests can be used as political tools and are thus open to manipulation unless due care is taken (in the form of ethical practices and other checks and balances) to prevent this situation (Shohamy 2001).

Furthermore, problems can arise as a result of insufficient awareness and inadequate knowledge of the social consequences of assessment (Ndaba 2005: 2). Ndaba (2005: 2) points out that "debate rarely addresses more fundamental issues concerning the social functions and outcomes of assessment". A potential danger is that politicians are sometimes in a position to control curricula and assessment criteria, and thus exercise an inordinate influence on what is taught and how it is taught. In this context, Marshall (2016) examines the changes made to testing in England and the extent to which these changes are politically driven. Marshall (2016: 14) discovered that teachers believed that the new tests had "little to do with educating pupils and much more to do with politics". Teachers complained strongly about frequent meddling by ministers who had little or no knowledge about pedagogical issues and general classroom practice. Marshall (2016: 1-2) observes that "Assessing pupil performance used to be an educational issue, and now it is not. Politics, even party-political politics, has become part of the assessment process".

In the case of high stakes examinations, these factors can present a very real threat. Although the target module is not as vulnerable as the extremely high stakes National Senior Certificate, the stakes remain high, as failure will delay the ultimate awarding of a degree, with all the associated social and financial consequences. The final examination takes place on a national, and even international level (as Unisa caters for students living outside South Africa), and forms part of a degree course. It is thus essential that the highest ethical standards are adhered to and that hidden agendas are not countenanced in the test and design assessment process. Care should be taken to prevent an over-emphasis by the university on pass rates or "throughput", with these rates being linked to merit bonuses for teaching staff. This could be equated to pressure on lecturers to pass students who have not necessarily met all the requirements of the module.

4.7 THE RATING SCALE 'MYTH': FURTHER OBSERVATIONS ON RATING SCALES

The researcher concurs with Spencer (1998: 132) that the concept of a perfect rating scale that fully reflects the students' ability and is totally reliable, is a myth. It is

possible only to strive for an approximation of the ideal of perfection. Spencer (1998: 132) acknowledges that all assessments are flawed and that “there is no validation process that is completely empirical, is completely impersonal and objective, and avoids the vagueness and uncertainty of human judgements altogether”. Huot (1990: 203) also cautions that “the test or observed score ... is a function of the true score and some component of error”. However, as Spencer (1998: 132) points out, “the fact that perfection is unattainable does not absolve the researcher from the responsibility of striving towards it”.

Literature on specific rating scales includes reviews of the International English Language Testing System (IELTS), Writing Test (Ysal 2010; Azizi & Majdeddin 2014; Ghamarian *et al.* 2014), the scale of Jacobs *et al.* (1981), and the TEEP Attribute Writing Scale (Weir, 1990). These scales, as well as the socio-cognitive framework posited by Shaw and Weir (2007), which deals in more detail with the validation process, are significant in the assessment process.

The scale used by the target group of this thesis (and also employed by language-based modules at Unisa), has been adapted from that proposed by Spencer (1998: 133-134) in a study of the response to assessment strategies by students of Practical English at Unisa (PEN 100-3) (Appendix B). This scale is a modification of the ESL Composition Profile of Jacobs *et al.* (1981). Spencer (1998: 133 - 134) observes that:

The *ESL Marking Profile*, used in the assessment of student writing... requires separate evaluation of form and content. This should promote formative assessment and invite a return to the creative chaos, the reworking stage that is so beneficial to student writing development.

Furthermore, the separation of content and form aims to prevent “mismatches between surface deficiencies and form” (Haswell & Wyche-Smith 1994: 228), a problem often encountered in the writing of second-language students. Another important feature of the scale is the distinction made in the language assessment section between surface errors (which do not affect meaning) and those which obscure meaning. This is a valuable distinction as it prevents the danger of focusing on form to the detriment of meaning, a problem encountered by Spencer (1997; 2005) and Spencer *et al.* (2005).

The elements of the scale are weighted according to their perceived relative degree of importance, and the total weight assigned is indicated by numerical ranges corresponding to four levels of competence, namely: “Excellent to Very Good”, “Good to Average”, “Fair to Poor”, and “Very Poor”.

Spencer (1998: 133) notes that the *ESL Composition Profile* has demonstrated reliability as a testing instrument. This claim is supported by research by Astika (1993) during which the essays of 201 subjects were rated by two or three markers, using the *ESL Composition Profile*. The correlation coefficient indicating inter-rater reliability proved to be 0.82, very similar to the findings of Jacobs *et al.* (1981) that showed an inter-rater reliability of 0.85.

Despite the strengths of the holistic scoring required when using the *ESL Composition Profile*, Spencer (1998: 133) acknowledges that holistic ratings have “major limitations”, including that:

- They cannot be used “beyond the population which generated them”.
- Training procedures related to holistic scoring can distort the process of scoring and reading and the rater’s ability to make sound choices (Huot, 1990: 201–2).

Spencer (1998: 133) also points out that fairness can be compromised because holistic scoring:

- rates “complex, multidimensional performances” by means of one-dimensional single numbers;
- gives no feedback to the student other than a single point on an “applause meter” without substantiating evidence;
- encourages the “dangerous assumption that there is a ‘true score’ for a piece of writing”;
- is based on “holistic global feelings” which are “the biggest enemy of thoughtful evaluation”;
- meets the need for ranking and evaluation.

In the case of the current target group, the holistic nature of the scale, especially in ODL with its limited interaction between markers, who come from diverse cultural and linguistic groups, might lead to subjectivity and different interpretations, especially in the light of the socio-cultural and multilingual factors involved. However, these can be addressed partially by means of training and moderation despite the relative lack of day-to-day communication between raters. For instance, in the current technological environment, more regular interaction can take place via electronic media such as email, Skype and WhatsApp. The training procedures criticised by Huot (1990: 201 - 2) could thus evolve into discussions leading to consensus, although time constraints present severe challenges in this respect (as demonstrated in Section 2.5). In the case of ENG1501, marking training is conducted online, and the results of moderation are emailed to the markers concerned.

Regarding the criticism that the simplicity of holistic scoring can be superficial and potentially obstruct “thoughtful evaluation” (Spencer 1998: 133), Cumming *et al.* (2002: 3) aver that this seeming simplicity hides the most pertinent strength of such a scale, which Lumley (2002: 24 - 248) describes as “reliance on the complex, richly informed judgements of skilled human raters to interpret the value and worth of students’ writing abilities”. This is dealt with in more detail in Chapters 6 and 7 in connection with feedback from markers and tutors.

A concern is that the assessment scale currently used is applied (with very minor adjustments) to various modules offered by the Department of English Studies at Unisa. For instance, while in all cases academic language proficiency is a common aim, the question that needs to be asked is whether the current scale is valid for all modules. The differences might be merely those of emphasis (for example, the scale’s mark weighting of content/organisation relative to language usage), but it is possible that these might be significant enough to affect validity and thus require a scale tailor-made to the specific outcomes of the module (Appendix C).

The extent and impact of subjectivity, as well as whether the limited communication between markers has substantive effects on rating, was borne in mind in the current study. The relative lack of interaction could affect the reliability of the scale, but it is

also important to interrogate its construct validity, given that the same scale is used for various modules and that it has been largely unchanged since 1998.

4.8 Conclusion

This chapter commenced with definitions of the term “validation”, followed by a discussion of the validation process with particular reference to the evidence-based argumentative approach as an effective and fair means of validation.

Language models and validation frameworks were then examined, as these formed a foundation for the development of an assessment instrument that meets the criterion of a “balanced scale that gives adequate feedback for both teachers and learners while being as practical as possible” (Hattingh 2009: 145). It was ascertained that the number of levels to be used would also be determined by the empirical means of pre-testing and piloting, as recommended by Weigle (2002: 127).

Factors that affect scores were described, including those directly related to learning, teaching and assessment, administrative and physical factors, and the impact of the assessment on institutions and society. The chapter concluded with a discussion of the perceived strengths and weaknesses of the current rating scale. These have been interrogated further in the empirical research described in the following chapters.

Chapter 5: Research Methodology

5.1 Introduction

This chapter contains a description of the method of research followed in this study to evaluate the existing rating scale used for *Foundations in English Literary Studies* (ENG1501) at Unisa and the subsequent modification or replacement of the existing scale. Firstly, the rationale for the chosen methodology is explained. This was based on the research questions addressed in the study. Details of the research instruments and data also are provided. A description of the research procedure is then given. This includes a discussion of the qualitative and quantitative processes employed at each stage. The chapter continues with a description of the ethical procedures that were followed by the researcher. The methods employed to address the primary and secondary research questions addressed by the research are then given as an overview of the research methodology. This overview is followed by a general summary of the process, and the topics covered in this chapter are summarised in the conclusion.

5.2 RESEARCH QUESTIONS

The purpose of this study was to investigate the validity of an existing assessment scale for academic writing on literary texts in a distance learning environment and, depending on the outcome of the investigation, to modify the existing scale with the aim of working towards an empirically-validated scale for assessing the assignments of the target group. The outcome was the validation of a rating scale appropriate to its purpose and context.

As noted in Chapter 1, the study addressed the following primary questions, based on the problem statement:

1. Is the existing assessment scale used for the *Foundations in English Literary Studies* (ENG1501) at Unisa valid in terms of the various aspects of validation and purposes (namely formative assessment, summative assessment, feedback)

This question gave rise to a further question, namely:

2. How can the existing scale be modified or replaced in order to produce a validated scale for assessing the essays of the target group?

These primary questions were supported by the following sub-questions:

- What do the results of the empirical research process reveal about the validity of the existing scale?
- What are the observations of the tutors and markers who use the existing scale to assess examinations and assignments for this module?
- What effect, if any, does the distance learning, multilingual and multicultural context have on the perceived and actual validity of the scale?
- What recommendations, principles and insights from other stakeholders can be employed when devising an improved scale?
- How can the modified or new rating scale be designed and tested to ensure optimum validity?

The methods employed to answer these questions have been explained in Section 5.3. General comments on the choice of design and methods are included in order to provide background to the subsequent discussion.

5.3 RATIONALE: RESEARCH DESIGN AND METHODS

An empirically-based procedure was adopted for this research in order to foster thoroughness and objectivity and to provide opportunities to re-check and cross-check the data obtained from the marking and planning processes. Quantitative processes were complemented by qualitative techniques.

The combination of quantitative and qualitative elements adopted for the current research project has been described as a mixed methods (MM) approach. Vorobel and Kim (2012: 255) point out that: “While quantitative research applies objective

measurement and statistical analysis of data in order to answer a research question... qualitative research involves an interpretive, naturalistic approach to the world". The researcher can implement qualitative methods by analysing sources such as notes, questionnaires, interviews and comments made by stakeholders. The process followed can be either parallel or sequential. A largely sequential process was followed in this study, whereby the "first phase of data collection can help to inform the second phase, or the second phase can be used to aid in the interpretation of data collected in the first phase" (Vorobel & Kim 2012: 255). This approach applied particularly to the sequencing of the marking, the statistical processes and the panel discussions, which culminated in the design and testing of the proposed new scales (Chapters 6 and 7). The value of combining research methods (as done for this project) is corroborated by Weir (2005: 15), and is also in accordance with the belief expressed by Bachman (2004: 6) and Kane (2017: 447 - 453) that both qualitative and quantitative approaches should be used to establish an assessment instrument's suitability in a particular context. This will allow for "an active search for meaning from the beginning, with the interpretation being elaborated and extended as data are collected" (Kane 2017: 453).

The quantitative elements in the current research study comprised the statistical evidence generated by the assignment results, while qualitative elements included the theoretical underpinning of the project, markers' comments, the information extracted from questionnaires, and comments from other stakeholders. The results of the statistical process were borne in mind when designing the questionnaires and, together with the results of the qualitative research, were considered during the panel discussions.

The theoretical underpinning encompassed a survey of literature examining the concept of validity, including different opinions on the relationship between, and relative importance of, the various types of validity. Theoretical models and frameworks were examined in order to establish a foundation for the development and validation of a fair and appropriate assessment scale. Rating scales were also examined with the aim of building a foundation for the development of an assessment instrument that met the criterion of a "balanced scale that gives adequate feedback for both teachers and learners while being as practical as possible" (Hattingh 2009: 145). With this aim in mind, the distance learning context in which assessment takes place was explored to

provide the background and context of the assessment and validation processes. The discussion included research on assessment in this environment, and particularly the impact of multilingual, cultural and socio-economic factors on the language assessment process in the ODL context. The challenges of teaching literature, particularly in the South African ODL environment, were also explored.

However, as explained in the description of the procedure employed for the current research project (Section 5.8), some activities, such as the continual updating of the theoretical underpinning, were carried out throughout the process and, in fact, the empirical findings pointed to additional areas of research, as visualised by Kane (2017: 447). Furthermore, at later stages in the project, such as during the panel discussions and testing phases, constant cross-referencing was made between the quantitative and qualitative data collection and analysis, although this remained largely sequential, following the pattern of one phase informing the next or the previous one (see Section 5.8).

A further aspect to consider is the assessment context or environment. Huot (1996: 161) emphasises that local standards should be considered during the validation of assessment procedures, including reading contexts and the background against which the design of assessments takes place. Huot (1996: 161) believes that, in order to ascertain the role of context in a specific assessment, qualitative procedures should be employed to complement the role of quantitative validation procedures, and to prevent the scale from having unfair social consequences (1996: 161 - 162). It is acknowledged that invalid assessment would affect the students' results and, by implication, result in negative social consequences, such as unwarranted failure and, ultimately, incomplete qualifications. This is one of the reasons why, in addition to the quantitative data generated by statistical procedures, the present research project includes qualitative information from stakeholders, such as markers and tutors. However, the heterogeneity of the target group presented limitations, as one cannot refer to a single set of local standards when examining the context of the study. The primary shared context is the multi-faceted one of ODL, as described in Chapter 2, which is characterised chiefly by geographical distance with its concomitant communication and logistical problems, and by demographic diversity, which can give rise to psychological and socio-economic "distance" between those involved in the teaching and learning process (Sections 5.6.1.

- 5.6.3). It is this ODL context that makes the present study unique, and it was the lack of commonality that gave rise to its greatest challenge.

Regarding language assessment, Cheng and Da Luca (2011: 104 - 105) point out that: “Contemporary validation practices in educational assessment and in language assessment rely on multiple frameworks to justify test validation and use” and that, whereas some of these concentrate on internal validity by examining psychometric processes, others broaden the scope by considering contextual factors and the related social consequences. In the present study, the research was guided by argument-based models (Kane 1992; 2004; Taylor 2002; Shaw & Weir 2007) in order to ensure the systematic collection of evidence at various stages of assessment. The models have been discussed in Chapter 4 of this thesis. Data were collected systematically for this project from a variety of participants such as panel members, online markers and tutors, external markers from other institutions, and Unisa Parow Regional Learning Centre staff. Cheng and Da Luca (2011: 105) note that a process of this kind comprises qualitative methodology to explore stakeholders’ “perspectives on their testing experiences in order to contribute broad validity evidence towards ongoing validation in... language testing”.

The advantage of MM is summed up by Tsushima (2015: 104), who observes that “MM research provides a holistic view of a research problem by combining quantitative and qualitative data in a single study”. This view is shared by several other researchers such as Creswell and Clark (2007), Greene (2007) and Teddlie and Tashakkori (2009). This MM approach was considered appropriate to the current research, as its field of study is a complex one, encompassing many facets, such as the pedagogical challenges arising from the ODL context, as well as the specific exigencies and constraints of the target module (see Chapter 2). It was for these reasons that a mixed research model was chosen to provide a “holistic picture” (Tsushima 2015: 104) and to ensure as much objectivity as possible.

No approach is without its disadvantages and, although MM research is valuable in the context of this project, a “practical and logistical issue” pointed out by Tsushima (2015: 105) is that, because the approach involves “several data collection stages, MM researchers need to develop multiple instruments”. This could be a particular problem in

the ODL context, where participants are often not known to one another and are frequently geographically separated. This separation potentially poses difficulties at many stages of the communication process between the stakeholders. In the current study, for instance, logistical issues occasionally caused problems in activities such as accessing the scripts, establishing contact with markers, obtaining the co-operation of Unisa online tutors, markers and moderators, and liaising with other support structures such as learning centres and administrative departments at Unisa. However, these problems were mitigated by the use of electronic and social media, such as emails, Skype, the MyUnisa website, SMS messages, Google and WhatsApp. These media provided opportunities for more rapid communication than was possible before their advent. Emails were particularly effective because they allowed for longer and more complex communication, and were particularly valuable in the exchange of marked and unmarked scripts. This prevented the onerous and time-consuming process of printing and photocopying (and, in some cases, retyping) the scripts selected for marking. WhatsApp and SMS messages served the purpose of quick communication, and were useful for conveying reminders, practical information and concise comments.

Tsushima (2015: 105) acknowledges that the research questions should be the chief determinants of research design, but cautions that “it is also sensible to plan a MM research study within a feasible timeline and financial constraints” (2015: 105). This statement was applicable to the current project, given the factors of geographical distance and the relative lack of interaction discussed previously (Chapter 2). Fortunately, the size of the final sample group ($n = 60$) ensured that, despite time and logistical constraints, the process was manageable in terms of the number of people involved and also the timeframe. Furthermore, provision was made for a timeframe that allowed for the potential logistical and administrative delays frequently experienced in the distance learning environment. Funding was provided by Unisa for the first four years of the study.

The timeline followed for this research had to be adjusted frequently to make allowance for delayed responses to emails as well as changing circumstances and commitments on the part of respondents and particularly markers who were all engaged in full- or part-time professional employment. Time also had to be allowed for the design of the new

scales and the subsequent testing thereof, as discussed in Chapter 7. The timeline was planned as follows:

- November 2016: Scripts accessed, downloaded and selected.
- November 2016 - March 2017: Markers briefed and marking commences, using current scale; panel members selected and contacted; comments on pilot results made; statistical data obtained and analysed; partial pilot project completed;
- June - September 2017: Marking of assignments (current scale). Allowance made for markers' full-time commitments such as marking end-of- semester examinations.
- September – October 2017: Comments from markers of this phase obtained and analysed.
- February 2018 – May 2018: Panel discussions.
- March – April 2018: Final statistical analysis of current scale.
- May 2018: Questionnaires on current scale sent to tutors and markers.
- February – May 2018: Meetings, ongoing emails and informal discussions with panel members by means of telephone and WhatsApp communication. This culminated in the design of the new models (Models 1 and 2)
- June – September 2018: Trial of Model 1 and 2.
- September 2018: Questionnaires on Models 1 and 2 sent to stakeholders; statistical data on these scales were calculated and analysed.
- September 2018 – May 2019: Qualitative and quantitative data obtained and analysed. Draft thesis written.
- January 2020 – Thesis submitted for examination.

5.4 DATA

The data were based on the first assignment of ENG1501 for the Second Semester of 2016. Students participating in the ENG1501 module were required to submit assignments on the topics prescribed by *Tutorial Letter 101* of 2016. These topics covered issues raised by the literature prescribed for the module, and tested the students' knowledge of the prescribed texts, as well as their insight into the various themes presented in the literary works. In addition, students were expected to show awareness

of the literary genres represented by the works prescribed for the module. Students were also assessed for language usage (Section 5.8). The assignment was chosen as the construct for this research for practical reasons, chiefly because of the accessibility of the scripts to the researcher, and the fact that the poem, *Small Passing* by Ingrid de Kok, on which the questions were based, was printed at the beginning of the assignment. This made it immediately accessible to the markers, who would otherwise have had to be provided with the prescribed work and a longer briefing and preparation period for would have been required for those markers not employed as Unisa tutors and/or markers. The assignment is shown in the table below.

Table 5.1: Module Assignment

SEMESTER 2

ASSIGNMENT 01: Seasons Come to Pass

Due date: 31 August 2016

Unique number: 674052

Read the poem below (pp. 254–5 in *Seasons Come to Pass*), and then answer the questions that follow. Each question on the poem should be answered in paragraph form (10–15 lines). Remember to quote from the poem to substantiate your answers.

Small Passing (Ingrid de Kok 1951 –)

For a woman whose baby died stillborn, and who was told by a man to stop mourning, ‘because the trials and horrors suffered daily by black women in this country are more significant than the loss of one white child’.

In this country you may not
suffer the death of your stillborn,
remember the last push into shadow and silence,
the useless wires and cords on your stomach,
the nurse’s face, the walls, the afterbirth in a basin. 5

Do not touch your breasts
still full of purpose.
Do not circle the house,
pack, unpack the small clothes.
Do not lie awake at night hearing 10
the doctor say ‘It was just as well’
and ‘You can have another.’

In this country you may not mourn small passings.

See: the newspaper boy in the rain 15
will sleep tonight in a doorway.
The woman in the bus line

may next month be on a train
to a place not her own.
The baby in the backyard now 20
will be sent to a tired aunt,
grow chubby, then lean,
return a stranger. Mandela's daughter tried to find her father
through the glass. She thought they'd let her touch him. 25

And this woman's hands are so heavy when she dusts
the photographs of other children
they fall to the floor and break.
Clumsy woman, she moves so slowly
as if in a funeral rite. 30

On the pavements the nannies meet.
These are legal gatherings.
They talk about everything, about home,
while the children play among them,
their skins like litmus, their bonnets clean. 35

2
Small wrist in the grave.
Baby no one carried live
between houses, among trees.
Child shot running,
stones in his pocket, 40
boy's swollen stomach full of hungry air.
Girls carrying babies
not much smaller than themselves.
Erosion. Soil washed down to the sea. 45

3
I think these mothers dream
headstones of the unborn.
Their mourning rises like a wall
no vine will cling to.
They will not tell you your suffering is white. 50
They will not say it is just as well.
They will not compete for the ashes of infants.
I think they may say to you:
Come with us to the place of mothers.
We will stroke your flat empty belly, 55
let you weep with us in the dark,
and arm you with one of our babies
to carry home on your back.

(Printed with kind permission of the poet. From Seasonal Fires: new and selected poems
by Ingrid de Kok, Umuzi 2006.)

Questions

1. The epigraph introduces a stark contrast between the 'small passing' and the everyday suffering of black South Africans. By referring to the first section (lines

- 1–35) of the poem, explain how the poet creates this contrast.
2. Identify the tone of the poem, and explain how it contributes to its meaning.
3. This poem is an example of free verse, which means that it has no set structure. However, the poet uses a number of poetic devices to create rhythm and form. Identify the main sound device the poet employs, and discuss its effect by referring to at least three examples from the poem.
4. Quote two similes from the poem, explain what they mean. Also consider how these similes develop meaning in the poem as a whole.
- 6 [sic]. Carefully consider lines 53–58, and explain how they contradict the rest of the poem.

Source: ENG1501 Tutorial Letter 101 of 2016, Unisa

A potential problem was that, although the assignment was to be assessed holistically, unlike the essay-type assignment (set for the other prescribed texts later in the semester), the poetry assignment consisted of a number of contextual questions. The move from an essay-type question to the short question format for the poetry assignment was because many students experienced great difficulty in writing an essay on poetry, possibly as a result of lack of training at school.¹¹ However, it was believed that, since the poetry assignment tested the given criteria as stated in the outcomes (Appendix C), it was possible to validate the scale by using evidence from the marking of the poetry assignment, provided clear briefing was given to markers. Students were required to write a paragraph of 10 to 15 lines on each question, and thus organisation, language use and coherence could be tested as well as interpretation and insight. It could also be argued that the grid should be tested for validity in order to ascertain whether and where it should be adapted to become more suitable for the full, current, assignment set.

Raw data were captured from the final sample of 60 scripts, chosen as described in the sampling method below (Section 5.5). The scripts were marked and the results were entered onto a spreadsheet for statistical and quantitative processing. The qualitative data were generated by comments requested from the markers of the scripts as well as answers to questionnaires sent to Unisa markers, moderators and online tutors.

¹¹ L2 learners studying English as an additional language (previously known as a Second Language) are not required to write essays in their Grade 12 school-leaving literature examination. All questions are contextual.

5.5 SAMPLING METHOD

The first step was to identify a representative sample of scripts received from the student target population of the *Foundations in English Literary Studies* (ENG1501) module for Assignment 1 of Semester 2, 2016. Participating students were selected from those submitting poetry assignments electronically as part of their formative assessment ($n = 3041$), which constituted the sample population for the study. An initial sample of 200 marked scripts was chosen at random. The researcher carried this out with the assistance of the module co-ordinator.

The 200 marked scripts were then divided into categories according to the 4 levels of the marking grid (see Section 5.8), after which random samples were drawn pro rata from each category level (as indicated in the tables below) to obtain a total sample of 60 scripts. The sample was selection according to the original mark allocated to the scripts by a Unisa marker. Thus, all four levels of the marking grid were represented by the smaller sample (Table 5.2). The data followed a Bell curve design.

Table 5.2: Distribution of sample scripts according to levels of the existing rating scale

Level	Number of scripts
Level 1 (100 – 76%)	2 (3.3% total)
Level 2 (75% – 56%)	16 (26.6% of total)
Level 3 (56% – 32%)	38 (63.3% of total)
Level 4 (30% – 0%)	4 (6.6% of total)
Total	60 (100% of total)*

* Selected according to the same ratio as the 200 randomly selected scripts

In effect, this sample was a stratified, random sample, where the strata were the levels of the existing rating scale. Stratification was used to ensure that the final sample was representative of the sample population with the same proportion of scripts within each level according to their existing scores. This was a consideration when harvesting the scripts. The sample made up 1.97% of the total 3041 students who submitted the assignment electronically. Although 1.97% is a small percentage of the total, consideration had to be given to the manageability of the sample and the availability of the data, bearing in mind that scripts were due to be archived soon after being downloaded. It is noted also that a validity study by Hattingh (2009: 89) used a much

smaller sample (n = 64) selected from 592 000 Grade 12 learners' scripts, making up 0.0108% of the total. Similarly, in the case of the current research, the aim of the process was to obtain a "representative sample of manageable size" (Rabiah 2010 : 417) and, in the case of the current research, the sample was deemed to sufficient, since it included examples of results at all four levels of the rating scale (as indicated in Section 5.8). These assignment scripts were captured electronically from the available data. Permission to use the scripts for research purposes was obtained from the Ethical Clearance Committee of Unisa (see Appendix A1).

5.6 SAMPLE POPULATION

The sample population comprised the total student body of first-year students of the Faculty of Arts and Humanities at the University of South Africa who were registered for *Foundations in English Literary Studies* (ENG1501). Students came from a diversity of demographic backgrounds and, in most cases English was not their home language. The information in the following sub-sections has been presented in Chapter 2 as evidence in support of the argument-based approach to the validating process and is presented here again as part of the sampling methodology for this study.

5.6.1 Sample composition by home language

Table 5.3 below shows the home languages of the sample group.

Table 5.3: Sample composition by language for Semester 1 and 2, 2016 (as at 27 October 2016)

Module Code	Home Language	Academic Year 2016	
		Semester 1	Semester 2
ENG1501	AFR/ENG	270	237
	AFRIKAANS	913	702
	ENGLISH	1763	1304
	FRENCH	6	3
	GERMAN	3	1
	GUJARATI	1	
	HEBREW	1	
	HINDI		2
	ISINDEBELE	115	79
	ISIXHOSA	628	424
	ISIZULU	3805	2911
	NDONGA		1
	NORTHERN SOTHO	570	432
	OTHER AFRICAN LANGUAGES	20	20
	OTHER FOREIGN LANGUAGES	4	3
	PORTUGUESE	5	1
	SESOTHO	287	303
	SETSWANA	394	342
	SHONA	80	48
	SISWATI	172	159
	SPANISH	1	
	TSHIVENDA	101	85
	Unknown	5	3
	XITSONGA	239	198
Total		9383	7258

Source: Reproduced with permission of Directorate: Information and Analysis (Unisa)

The figures in Table 5.3 show that, although a fairly sizeable group (1763 and 1304 for Semester 1 and 2 respectively) claimed English as their home language, this number is eclipsed by other groups. The largest group is Zulu home language speakers with 3805 students (40.5%) and 2911 (40.1%) for semester 1 and 2 respectively. The English Home Language group was in the minority with 18.89% and 17.96 % of the totals for

Semester 1 and 2 respectively, compared with the total number of speakers of other languages being 7620 (81.21%) for Semester 1 and 5954 (82%) for Semester 2.

5.6.2 Sample composition by province and residential regional office

Table 5.4 shows the geographical distribution of the target group according to Province and Residential Regional Office.

Table 5.4: Sample composition by province and residential regional office for Semester 1 and 2, 2016 (as at 27 October 2016)

Module Code	Residential Province	Residential Regional Office	Academic Year 2016	
			Semester 1	Semester 2
ENG1501	EASTERN CAPE	East London	87	73
		Mthatha (Umtata)	169	103
		Port Elizabeth	149	119
		Wildcoast (Mbizana)	20	12
	EASTERN CAPE Total		425	307
	FREE STATE	Bloemfontein	83	70
		Kroonstad	72	119
	FREE STATE Total		155	189
	GAUTENG	Ekurhuleni (Benoni)	837	654
		Florida	439	338
		Johannesburg	616	429
		Pretoria / Sunnyside	693	556
		Vaal Triangle	146	129
	GAUTENG Total		2731	2106
	KWAZULU NATAL	Durban	2370	1673
		Newcastle	277	252
		Pietermaritzburg	651	466
		Richards Bay	523	480
		Wildcoast (Mbizana)	163	94
	KWAZULU NATAL Total		3984	2965
	LIMPOPO	Giyani	126	109
		Makhado (Louis Trichardt)	36	36
		Polokwane	198	153
		Tlhabane (Rustenburg)	5	2
	LIMPOPO Total		365	300
	MPUMALANGA	Middelburg (Mpumalanga)	307	228
		Nelspruit	161	164
	MPUMALANGA Total		468	392

	NORTH WEST	Kimberley	4	2
		Mafikeng	91	93
		Potchefstroom	55	63
Module Code	Residential Province	Residential Regional Office	Semester 1	Semester 2
		Tlhabane (Rustenburg)	191	161
	NORTH WEST Total		341	319
	NORTHERN CAPE	Kimberley	81	69
	NORTHERN CAPE Total		81	69
	Unknown	Unknown	135	87
	Unknown Total		135	87
	WESTERN CAPE	George	85	83
		Parow	613	441
	WESTERN CAPE Total		698	524
Total			9383	7258

Source: Reproduced with permission of Directorate: Information and Analysis (Unisa).

Table 5.4 shows the geographical spread of the student population of ENG1501 with the greatest concentration being in KwaZulu Natal with 3984 (42.49%) and 2965 (40.8%) of the total for Semester 1 and 2 respectively, particularly in the Durban area with 2370 students (25.2%) and 1673 (23%) for Semester 1 and 2 respectively. This information correlates with the data on Home Language, as Zulu is the dominant Home Language in this area.

5.6.2 Sample composition by race and gender

Table 5.5 shows a breakdown of the number of students in the target group according to race and gender.

**Table 5.5: Sample composition by race and gender for Semester 1 and 2, 2016
(as at 27 October 2016)**

			Academic Year 2016	
Module Code	Race	Gender	Semester 1	Semester 2
ENG1501	African	Female	5565	4325
		Male	1167	893
	African Total		6732	5218
	Coloured	Female	414	321
		Male	75	53
	Coloured Total		489	374
	Indian	Female	460	354
		Male	67	49
	Indian Total		527	403
	Unknown	Female	30	16
		Male	5	9
Module Code	Race	Gender	Semester 1	Semester 2
	Unknown Total		35	25
	White	Female	1287	1010
		Male	313	228
	White Total		1600	1238
Total			9383	7258

Source: Reproduced with permission of Directorate: Information and Analysis (Unisa)

As shown in Table 5.5, African students predominate with registrations of this group totalling 6732 for Semester 1 and 5218 for Semester 2. This constitutes 71.7% and 71.8% of the total registrations for Semester 1 and Semester 2 respectively. White students comprised the second highest group, totalling 1600 (17% of total) for Semester 1 and 1238 (17% of total) for Semester 2. There were more female than male students in all groups with a total of 7756 (82.6%) and 6026 (83%) for Semester 1 and 2 respectively. The largest group represented was female African students (5565 or 59.3% of the total for Semester1 and 4325 or 59.5% for Semester 2).

5.7 MEASURING INSTRUMENTS

Instruments used to gather data included structured questionnaires and comments provided by the markers of the sample scripts.

The aim of the questionnaires was to obtain information about the needs, views and concerns of stakeholders. Structured questionnaires were distributed to:

- markers (after the measuring processes)
- online tutors
- panel members
- markers after the trialing process.

The main purpose of the questionnaires was to find answers to the research questions. All questionnaires included closed-ended questions (for specific information) as well as open-ended questions (requiring longer answers to supplement the information and provide opportunities to expand on the answers to the closed-ended questions by expressing opinions and making suggestions). The latter were included to elicit rich data.

The questionnaires that were sent by email to all Unisa markers and tutors of ENG1501 included questions on whether the assessment scale should make provision for the Open Distance Learning (ODL) context, and, if so, the form that this provision should take. Contact details for all Unisa markers and e-tutors for ENG1501 were supplied by the co-ordinator of the module as well as the co-ordinator of the e-tutor programme. Permission had been obtained from the Ethics Committee (Appendix A1) and participants were asked to sign the consent form (Appendices A2- A4). Furthermore, the complex ODL context was taken into account during the panel sessions, which culminated in the design and piloting of a new rating scale. The questionnaires dealing with the existing scale addressed the important issue of the number of levels provided for in this scale, as well as the division of the scale into separate sections for language and content/organisation. A further questionnaire, completed by panel members and markers after the testing of the new or modified scales, addressed similar issues, this time in respect of the proposed scales.

Table 5.6 below shows a copy of the questionnaire distributed to Unisa markers. The tutors' questionnaire did not include Question 6 as it related to marking. In retrospect, the same questionnaire could have been sent to both groups with the instruction to tutors

not to answer Question 6. However, despite this proviso, it was relatively easy for the researcher to analyse and summarise the information.

Table 5.6: Markers' questionnaire

<p>Markers' questionnaire ENG1501</p> <p>Please assist the researcher by completing the following questionnaire. Please note that you may answer anonymously if you wish.</p> <p>Name (optional) _____</p> <p>A. Background information.</p> <p>Please provide the following information by placing a cross (x) in the applicable box(es).</p> <p>What is your current position on the ENG1501 team?</p> <table style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 20%; border: 1px solid black; height: 20px;"></td> <td style="width: 20%; border: 1px solid black; height: 20px;"></td> <td style="border: 1px solid black; padding: 2px 10px;">Marker</td> </tr> <tr> <td style="border: 1px solid black; height: 20px;"></td> <td style="border: 1px solid black; height: 20px;"></td> <td style="border: 1px solid black; padding: 2px 10px;">Moderator</td> </tr> <tr> <td style="border: 1px solid black; height: 20px;"></td> <td style="border: 1px solid black; height: 20px;"></td> <td style="border: 1px solid black; padding: 2px 10px;">Examiner</td> </tr> <tr> <td style="border: 1px solid black; height: 20px;"></td> <td style="border: 1px solid black; height: 20px;"></td> <td style="border: 1px solid black; padding: 2px 10px;">Other</td> </tr> </table> <p>If 'other', please specify</p> <p>B. FEEDBACK ON THE CURRENT RATING SCALE</p> <p>Please indicate your answer by placing a cross (x) in the relevant box.</p> <p>1. Do you think that the scale adequately assesses the construct of the module as stated in the outcomes (i.e. does it test what it is supposed to test)?</p> <table style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 20%; border: 1px solid black; padding: 2px 10px;">Yes</td> <td style="width: 20%; border: 1px solid black; height: 20px;"></td> </tr> <tr> <td style="border: 1px solid black; padding: 2px 10px;">No</td> <td style="border: 1px solid black; height: 20px;"></td> </tr> <tr> <td style="border: 1px solid black; padding: 2px 10px;">Partially</td> <td style="border: 1px solid black; height: 20px;"></td> </tr> </table> <p>Comments and reasons</p> <p>2. In your opinion, is the distinction between the band levels clear?</p> <table style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 20%; border: 1px solid black; padding: 2px 10px;">Yes</td> <td style="width: 20%; border: 1px solid black; height: 20px;"></td> </tr> <tr> <td style="border: 1px solid black; padding: 2px 10px;">No</td> <td style="border: 1px solid black; height: 20px;"></td> </tr> <tr> <td style="border: 1px solid black; padding: 2px 10px;">Sometimes</td> <td style="border: 1px solid black; height: 20px;"></td> </tr> </table> <p>Comment, with examples</p> <p>3. Are there sufficient levels?</p>				Marker			Moderator			Examiner			Other	Yes		No		Partially		Yes		No		Sometimes	
		Marker																							
		Moderator																							
		Examiner																							
		Other																							
Yes																									
No																									
Partially																									
Yes																									
No																									
Sometimes																									

Yes	<input type="checkbox"/>
No, too many	<input type="checkbox"/>
No, too few	<input type="checkbox"/>

Give reasons for your answer

4. Do you agree with the present 50/50 weighting of marks between organisation/ content and language?

Yes	<input type="checkbox"/>
No	<input type="checkbox"/>

Reasons and comments (if you answered 'no', please give suggested alternative weighting).

5. Are there any features of the scale that you think are open to misinterpretation or subjectivity?

Yes	<input type="checkbox"/>
No	<input type="checkbox"/>

If 'yes', give examples and comments

6. What is your preferred approach when you mark ENG1501 assignments/examination answers?

I adhere strictly to the rating scale	<input type="checkbox"/>
I use the scale as a guideline, but use my own discretion	<input type="checkbox"/>
I ignore the scale and give an impression mark	<input type="checkbox"/>
It depends on circumstances (specify below)	<input type="checkbox"/>

Reasons and comments

7. Should the scale be designed to take the multicultural and multilinguistic distance learning target market into account?

Yes	<input type="checkbox"/>
No	<input type="checkbox"/>

Discuss and give reasons for your answer.

8. If you answered Number 7 in the affirmative, do you think that the scale adequately reflects the distance learning context? If not, how can it be amended?

Yes	<input type="checkbox"/>
No	<input type="checkbox"/>

Comments and suggestions

General observations

9. If you could make one change to the current scale, what would that change be?
10. What is the main feature that you would like the revised scale to retain?

11. Do you have any further suggestions or comments about the rating scale? If so, please mention these in the space provided below. You are also welcome to contact me at maxibob@telkomsa.net
Many thanks for your assistance. It is greatly appreciated.

Questionnaires were distributed to the markers of the two alternative rating scales (named Model 1 and Model 2 respectively) that were designed by the panel and these questionnaires were sent to the panel members themselves, after they had participated in a marking session to test the scales. Two identical questionnaires were provided, one for Model 1 and the other for Model 2. These were clearly marked to avoid confusion.

Table 5.7: Questionnaire for feedback on Models 1 and 2

Questionnaire ENG1501: Revised rating scale feedback	
Please assist the researcher by completing the following questionnaire. Please note that you may answer anonymously if you wish.	
Name (optional) _____	
FEEDBACK ON THE REVISED RATING SCALE	
Please indicate your answer by placing a cross (x) in the relevant box.	
1. Do you think that the revised scale assesses the construct of the module better than the current scale?	
Yes	<input type="checkbox"/>
No	<input type="checkbox"/>
Partially	<input type="checkbox"/>
Comments and reasons	
2. In your opinion, is the distinction between the band levels clearer than in the scale currently in use?	
Yes	<input type="checkbox"/>
No	<input type="checkbox"/>
Sometimes	<input type="checkbox"/>
Comment, with examples	
3. Do you think that the increased number of levels is sufficient?	
Yes	<input type="checkbox"/>
No, too many	<input type="checkbox"/>
No, too few	<input type="checkbox"/>
Give reasons	for your answer

4. Do you believe that the weighting of marks between organisation/ content and language on the grid produces a fairer score than the scale currently in use?	
Yes	<input type="checkbox"/>
No	<input type="checkbox"/>
Reasons and comments (if you answered 'no', please give suggested alternative weighting).	
5. In your opinion, are the criteria clearer in comparison to the current scale?	
Yes	<input type="checkbox"/>
No	<input type="checkbox"/>
Sometimes	<input type="checkbox"/>
Unsure	<input type="checkbox"/>
Reasons and comments (NB Please specify if you answered 'No' or 'Sometimes')	
5. Are there any features of the scale that you think are open to misinterpretation or subjectivity?	
Yes	<input type="checkbox"/>
No	<input type="checkbox"/>
If 'yes', give examples and comments	
7. Does the new scale take the multicultural and multilinguistic distance learning target market into account?	
Yes	<input type="checkbox"/>
No	<input type="checkbox"/>
Discuss and give reasons for your answer	
8. If you answered Number 7 in the negative, how can the scale be amended to reflect the distance learning context adequately?	
Comments and suggestions	
General observations	
9. If you could make one change to the revised scale (Model 1), what would that change be?	
10. What is the main feature that you would like the revised scale (Model x) to retain?	
11. Do you have any further suggestions or comments about the rating scale? If so, please mention these in the space provided below. You are also welcome to contact me at maxibob@telkomsa.net	
Thank you for your assistance. It is greatly appreciated.	

In summary, recommendations included comments from markers, tutors and panel members. These were extracted from the feedback from markers during and after the marking of the scripts, and from the results of questionnaires obtained from Unisa markers and tutors, as well as from panel members. The results emanating from this feedback were discussed further by the panel members and incorporated into the improved scale. Details of this feedback are provided in Chapters 6 and 7.

5.8 RESEARCH PROCEDURE

As described in Section 5.6.1, the student population investigated in this study comprised students who were registered for the *Foundations in English Literary Studies* (ENG1501) module at the University of South Africa, who submitted the first assignment set for the second semester of 2016 as part of their formative assessment (n = 3041). The target assignment was based on a prescribed poem and thus tested the students' knowledge of, and insight into, this particular genre, as well as their ability to express themselves in academic English.

The assignments used in this study had been submitted for assessment as part of the compulsory written work required by the ENG1501 module. Since the students were writing these assignments in a distance learning context, these assignments were written in a variety of venues, depending on the students' circumstances. The sample chosen by the researcher consisted of marked scripts that were still accessible on the database (Assignment 1 of the Second Semester 2016). The process followed was that, once the scripts were received from the students, they were stored electronically and distributed to the marking panel by the primary lecturer, who was an experienced Unisa lecturer assigned to co-ordinate the module. Scripts are marked by Unisa markers and returned to the student electronically. Shortly thereafter, the scripts are archived and are only available to certain staff members via the myUnisa system. The researcher thus had to make use of this 'window period' to access and download the scripts. Permission to access the scripts had been obtained from Unisa (Appendix A1).

Once the selection of 60 scripts had been made (as described in Section 6.2), the students' names, allocated marks and online marker's comments were removed and

randomly numbered copies of the selected assignments were sent to markers who allocated marks to them, for content and language use, according to the existing rating scale. Ten expert and experienced markers, as described in 5.8.2.1 below, participated in this process. The markers were familiarised with the current rating scale by means of a one-day briefing session and this ensured that they had a clear understanding of the assessment context and of the construct being assessed. The current rating scale grid follows

Table 5.8: Marking grid for ENG1501

Marking Grid (Content/Organisation – 25 , Vocabulary, Language Usage, Mechanics – 25) ENG1501		
Mark out of 25 for Content/Organisation		
SCORE	LEVEL	CRITERIA
25-19 (100%-76%)	1 EXCELLENT TO VERY GOOD	Content: focused on assigned topic, thoroughly developed, clearly demonstrating the skills required by the NQF criteria (e.g. familiarity with - recognising and recalling - the subject matter; understanding it; application of this information; analysis, for instance of relationships; evaluation, for example critiquing different approaches) Organisation: generating a piece of writing (such as an essay) with ideas clearly stated, succinct, well-organised, logically sequenced, cohesive, well-supported
18-14 (75%-56%)	2 GOOD TO AVERAGE	Content: fairly sound demonstration of skills, mostly relevant to topic, lacks detail Organisation: loosely organised, logical but incomplete sequencing and signposting
13-8 (54%-32%)	3 FAIR TO SHAKY: AT RISK	Content: not enough substance or relevance, insufficient support for ideas Organisation: ideas confused or disconnected, not enough logical sequencing or development, little signposting
7-0 (30%-0%)	4 VERY SHAKY	Content: not pertinent or not enough material to evaluate Organisation: does not communicate, no organisation or not enough material to evaluate

Mark out of 25 for Form (Vocabulary, Language Usage, Mechanics)		
SCORE	LEVEL	CRITERIA
25-19 (100%-76%)	1 EXCELLENT TO VERY GOOD	Vocabulary: sophisticated range, effective word/idiom choice, mastery of word form, appropriate register Language usage: effective complex constructions, few language problems (agreement, tense, number, word order, articles, pronouns, prepositions) Mechanics: mastery of presentation: neatness, spelling, punctuation, capitalisation, paragraphing and essay structure; meticulous and consistent referencing of sources used
18-14 (75%-56%)	2 GOOD TO AVERAGE	Vocabulary: satisfactory range, occasional issues of word choice, idiom, form, usage, but meaning not obscured Language usage: effective simple constructions, minor problems in complex constructions, several language issues but meaning seldom obscured Mechanics: occasional problems in mechanics

13-8 (54%-32%)	3 FAIR TO SHAKY: AT RISK	Vocabulary: small range, frequent issues of word/idiom, choice, usage Language usage: major problems in simple/complex constructions, frequent language issues including sentence construction problems, meaning confused or obscured Mechanics: frequent problems with mechanics, untidy handwriting, meaning confused or obscured
7-0 (30%-0%)	4 VERY SHAKY	Vocabulary: essentially translation from mother tongue, little knowledge of English vocabulary, idioms, word forms, or not enough material to evaluate Language usage: virtually no mastery of sentence construction, dominated by problems, does not communicate, or not enough material to evaluate Mechanics: no mastery of conventions, dominated by problems in mechanics, handwriting illegible, or not enough material to evaluate

5.8.1 Pilot study

Before the bulk of the scripts were marked, it was decided to test the main research procedures by means of a pilot study. Although this was curtailed by time constraints and funding difficulties, the pilot yielded guidelines for the main study. The findings have been presented in Chapter 6, which includes an account of the procedure and findings of this stage of the research. The present chapter is confined to the methodology and research stages of the project.

The purpose of the pilot study was to test the planned procedure and instruments. The steps followed reflected those of the main study as follows.

5.8.1.1 Collecting data and marking scripts

In the pilot study, 20 randomly selected scripts were marked by four markers who had been briefed and trained by the researcher. All four markers were qualified educationists and had experience of teaching English in a tertiary and secondary context. Furthermore, two of the markers had tutored and marked the ENG1501 module for Unisa.

The 20 scripts were chosen from the 60 selected for the project, and, since no problems with this step of the procedure were identified, the results of the marking were added to the later, larger study. The 20 scripts were each marked by all four markers and the results were recorded on a spreadsheet. Marks were allocated for content and language use and a total mark was indicated in each case.

5.8.1.2 Analysis of data and piloting of instruments

The data were then statistically analysed, using the Rasch programme described in Section 5.8.2.2, in order to test scoring validity (reliability), as defined in Chapter 3. The three features examined were content, language and the total mark. Unfortunately, results for Marker 1, the original Unisa marker, were available only for the total. This was because of the time constraints present in this pressurised modular and distance learning environment (described in Chapter 2).

The group of markers was requested to submit comments on the marking process and to pilot the questionnaires designed for the tutors and markers. This led to the modification and refining of the questionnaires, particularly in respect of questionnaire design and the range of topics covered. The process and findings have been discussed in Chapter 6.

5.8.1.3 Modification of existing scale or development of new scale – piloting of procedure of the panel discussions

This section of the pilot test was truncated owing to time constraints and logistical problems, but contact was made with the five experts who later formed the panel for the main study. During this contact, the pilot process and findings were briefly examined and the comments of the four markers were considered. Various options regarding alternative rating scales were discussed and arrangements made for future meetings during the main research study.

The panel of five experts included experienced educators and examiners of English at Unisa and other institutions.

The composition of the panel was as follows:

- Two members with PhD degrees and experience at secondary and tertiary institutions (including Unisa). One member was responsible for founding the De Beers English Olympiad and serves on the Olympiad co-ordinating committee. He has authored and co-authored textbooks for English Home Language and English First Additional Language as well as publications dealing with SAE

(South African English). The other member was a semi-retired senior lecturer at Unisa who is still involved with the marking of Unisa assignments and examinations.

- One member (the researcher) with a Master's degree and experience of teaching English Home and First Additional Language at secondary and tertiary levels, including at a TVET College, the Cape Peninsula University of Technology (CPUT) and Unisa (marking, tutoring at the Parow Regional Centre, online tutoring and writing material for English courses). This member has co-authored textbooks for English First Additional Language (Grade 12) and English Level 3 and 4 (TVET Colleges). She has been the Chief Examiner of the De Beers English Olympiad (1991; 2017) and has served on its selection committee.
- Two members with Honours degrees and experience at secondary and tertiary levels. Experience also included provincial co-ordination of Grade 9 literacy and numeracy testing, co-authoring of English Home Language and First Additional Language textbooks, and marking and examining the De Beers English Olympiad. One member was the Provincial Moderator for Senior Certificate (Grade 12) external examinations (English Home Language)

All panel members had experience in teaching English as L1 and as L2, and had all taught in multicultural and multilingual South African environments. Two panel members were L2 speakers of English but at this stage of the research educators who were L1 speakers of indigenous languages, other than Afrikaans, were unavailable. At a later stage (after the initial workshops), two colleagues, speakers of Xhosa and Zulu respectively, were consulted, and an informal group formed around emails and telephone conversations. Although the initial panel represented only two ethnic groups (white $n = 4$ and Coloured $n = 1$), it was deemed that their skill and experience was sufficient to ensure balanced and expert input.

5.8.2 Main study

Once the pilot project had been concluded, the main study commenced, following a similar procedure as that adopted for the pilot study. The steps taken were as follows.

5.8.2.1 Collecting data and marking scripts

As described in Section 5.8.1.1, after the 60 scripts had been selected, the students' names, marks and online markers' comments were removed and randomly numbered copies of the selected scripts were sent to markers who allocated marks to them, using the existing rating scale. Ten qualified and experienced markers participated in this process.¹²

The group of assessors included markers and tutors of ENG1501 as well as lecturers and examiners of similar modules at other tertiary institutions. Markers included two markers with PhD degrees; two markers with Master's degrees; three markers with Honours Degrees and three markers with other postgraduate qualifications (such as postgraduate Secondary Teacher's Diplomas as well as diplomas in Remedial Education). The latter group included a marker with experience of Matriculation examining, moderating, and materials writing for Grade 12 learners (L2), and two markers who had experience as examiners and moderators of the De Beers English Olympiad, a national examination for senior High School students (L1 and L2). All assessors were experienced educators, with a mean of 30 years' experience. Most markers ($n = 7$) were L1 speakers of English and three claimed both English and Afrikaans as their "home" languages. Two markers were fluent or fairly fluent in Xhosa. The ethnic grouping included White ($n = 8$), Coloured ($n = 1$) and Asian ($n = 1$). Once again, as mentioned in connection with the composition of the panel (Section 5.8.1.3), other racial groups were not represented as none were available. However, all markers had taught in a multicultural and multilingual environment, and had extensive experience in marking English Literature examinations and assignments.

In a briefing session, the marking panel had been familiarised with the existing rating scale and, consequently, could share an understanding of the assessment context and of the construct to be assessed. Training and briefing were reinforced by means of electronic media, mainly in the form of emails between the researcher and the markers for the purpose of clarifying issues arising during the course of marking, although the

¹² Following statistical advice, the initial group of 5 markers was expanded to a larger group of 10 in order to ensure greater reliability.

researcher did not attempt to influence the final mark awarded, except in the case of one marker, whose marks were initially erratic and who required a further briefing session. Subsequently, the marker produced consistent marks, although these remained lower than the results of the other markers. However, it was decided to retain this marker once the initial moderation and further briefing had taken place, as marks were consistent and the criteria were being followed.

In order to ensure a range of marking data, and thus an accurate reliability result, each marker was expected to rate at least 30 assignments, and each essay was rated by at least five of the 10 markers (following Hattingh 2009: 169). Marks were assigned for language use and content, as indicated by the scale (Appendix B), and markers were also required to allocate a combined score.

5.8.2.2 Analysis of data

After the marking process, the data were statistically analysed. The purpose of the analysis was to examine:

- the scoring consistency among the markers;
- the degree to which the sample of essays represented the full range of student competence on the scale;
- the accuracy of the levels at which essays were benchmarked by the raters;
- rater bias;
- the degree to which the rating instrument represents the construct of assessment.

The Rasch analysis employed by this research is a valuable, multi-faceted, procedure that provides “conclusive documentation of the many ways in which rater behaviour can vary, as well as to identify some of the kinds of measures (such as training and multiple rating) that can be taken to assist in managing this variation” (Lumley & Brown 2005: 830). The Rasch reliability index uses data of various facets of the marking process (such as learner ability, rater characteristics and item difficulty) in order to indicate the relationship between these facets, predict the student’s possible score, and investigate the differences between levels of scores assigned by different raters. Data can be used to demonstrate

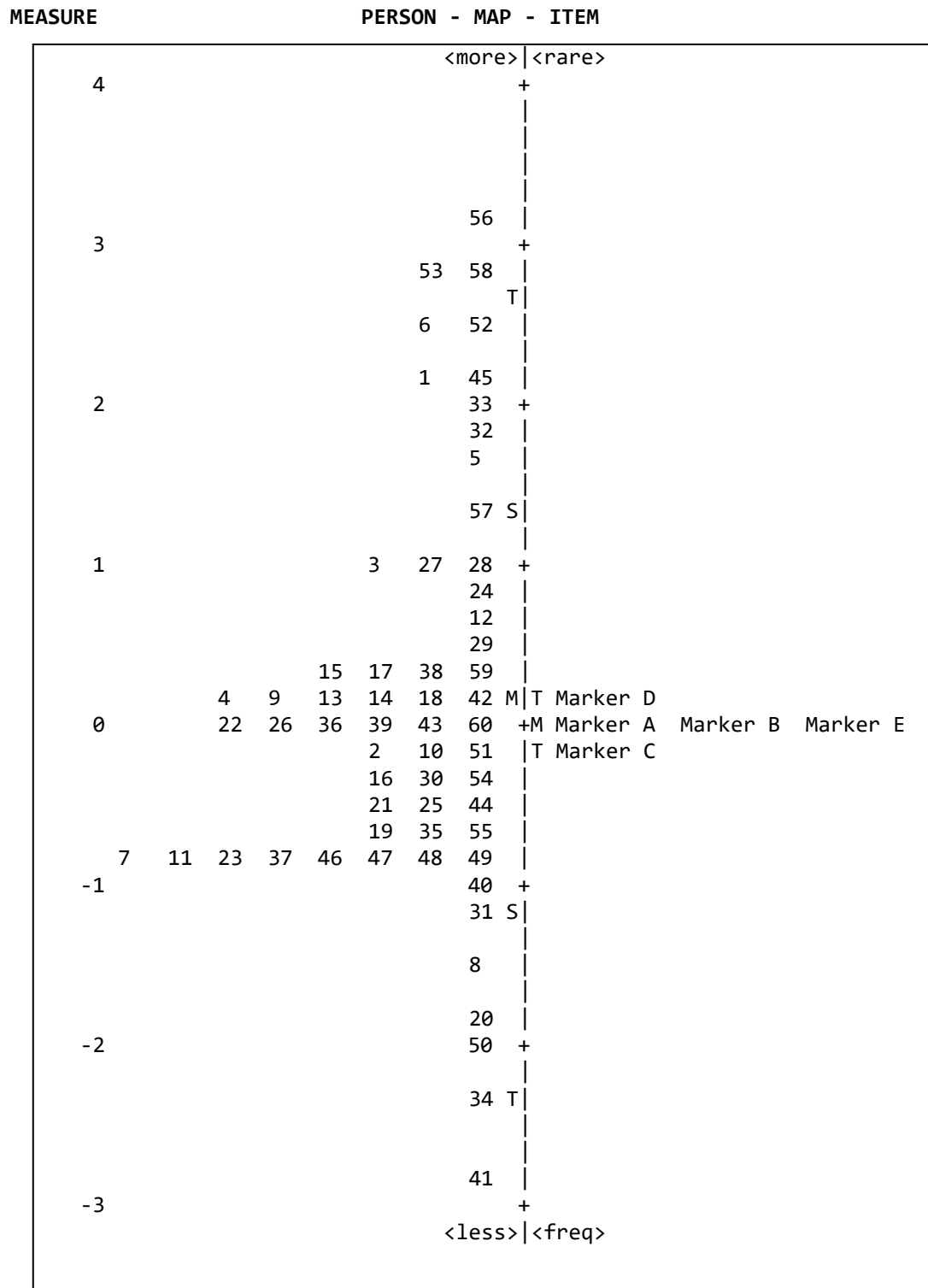
- item difficulty and rater bias towards any of the features or criteria of the rating scale;
- the degree to which the features measure the same construct;
- the degree to which the features indicated on the rating scale reflect the construct being assessed.

The advantage of using the multi-faceted Rasch measurement procedure is that rater characteristics, not merely raw scores, are taken into account (McNamara 1996: 118, Lumley & Brown 2005: 830). McNamara (1996: 133) points out that the Rasch model makes it possible “to bring all the facets together into a single relationship, expressed in terms of the effect they are likely to have on a candidate's chance of success”. McNamara (1996: 133) elucidates this statement by explaining that the Rasch model enables researchers to “see precisely what sort of challenge the candidate was facing on that criterion with [regard to] that rater, and are accordingly able to interpret the actual rating given”.

Similarly, Lumley and Brown (2005: 830) observe that: "Perhaps the most significant achievement of Rasch analysis has been conclusive documentation of the many ways in which rater behaviour can vary, as well as to identify some of the kinds of measures (such as training and multiple rating) that can be taken to assist in managing this variation". As McNamara (1996: 9) points out, the multi-faceted approach adopted by the Rasch analysis is valuable because, firstly, it accounts for inter-rater variance and, secondly, it provides an accurate indication of students' abilities. The analysis uses information on the various facets of the data matrix to predict the interaction between these facets (such as student ability, rater characteristics and item difficulty) and the likely score for the combination of the facets. This is achieved by the FACETS model provided by the Winsteps Rasch programme described in the following paragraph. McNamara (1996: 133) explains that “the model states that the likelihood of a particular rating on an item for a particular candidate can be predicted mathematically from the ability of the candidate, the difficulty of the item and the severity of the rater”.

The mathematical procedure used by Rasch (FACETS) is named the maximum likelihood estimation (McNamara 1996: 161) or “calibration”. This is recursive, repeating itself until the required level of accuracy of prediction is achieved. FACETS then provides a number of reports, notably a vertical ruler report providing information about the interaction between the different facets under investigation, e.g. test taker ability, task difficulty and rater characteristics. FACETS also indicates reliability indices and reports on unexpected results (Linacre, 2006b). Facet measurement reports provide detailed information on the interaction between different facets in the form of “fit” statistics. Fit statistics indicate the degree to which the data fit the Rasch model. If the pattern of the data does not fit the probabilistic model, the data are identified as misfitting (i.e. overfit). Values greater than +2 and smaller than -2 signify particular bias or misfit (McNamara 1996: 143, 173-174; Bachman 2004: 147; Hattingh 2009: 171).

Fit statistics can be used to identify unsuitable test items or assessment criteria or to determine the degree to which criteria and/or salient features addressed in a scale measure the same construct (Lumley & Brown 2005: 830). Misfitting items might indicate that criteria are poorly written or do not discriminate well between learners with different abilities. Alternatively, they could indicate that the item or criterion is good in itself, but that there is a problem of interpretation. The misfitting scores might be caused by a marker who does not measure the same ability or construct as the other markers who assess the assignment or test using the scale items in the test/scale (McNamara, 1996: 175). Raters identified as misfitting can be examined and modified or deleted from the test or scale if necessary. Figure 5.1 shows an example of a measurement report. In the case of the current research, the term “person” refers to the test-takers (represented by assignment numbers 1 to 60 on the left of the map) and the term “item” indicates the markers (represented by alphabetical letters A to C on the right). The report gives estimation, expressed in logits, of the test-takers’ ability. Test-takers placed above the 0 logit mark are more likely to answer a particular question correctly (i.e. higher ability), while those below the 0 mark are less likely (i.e. lower ability) (McNamara, 1996: 136).



Source: Rasch Winsteps Programme (FACETS)

Figure 5.1: Example fit statistics report for new scale trial Model 2

In the example above, raters are grouped in a cloud-like pattern around the 0 logit mark. This indicates consistency among the markers. On the other hand, student ability as seen in the scripts (represented as “persons” on the left-hand side of the figure) shows a wide

range, from 3 to -3 logits. This could be attributed to the diverse target group (as described in Sections 5.6.1.–5.6.3). The variance could also be a reflection on the assignment difficulty in relation to the scripts and not necessarily an indication of misfitting features of the rating scale, since the scale should indicate criteria pertinent to the module outcomes governing the assignment. This issue is explored in Chapters 2, 6 and 7 of this thesis and the map in Figure 5.1 serves merely as an example here.

Table 5.9 below is an example of the Rasch reliability index showing the reliability of the data. The Rasch reliability index is scaled from 0 to 1, with values closer to 1 indicating a good reliability rating. In Table 5.9, “person” denotes the test taker, represented by the script in this context; the Mean is the totals of the 5 markers for the 60 scripts; PSD indicates the Population Standard Deviation of the 60 script totals.

Person reliability indicates the reproducibility of the measure or test. High reliability of persons means that it is highly probable that persons with estimated high or low measures actually do have higher or lower abilities than other participants. The Item reliability, in the summary of 5 measured items, indicates the marker reliability.

The Cronbach Alpha score indicates the conventional test reliability index. It indicates the degree of relation between scale items. Scores vary from 0 to 1, with scores of 0.75 or higher considered to be consistent. Thus, the higher the alpha score, the more similar the items are likely to be. A high coefficient value demonstrates consistency between raters for the same sample of scripts.

Table 5.9: Example reliability index for data new scale trial Model 2

Input: 60 Persons 5 Item. Reported: 60 Person 5 Item 73 Cats Winsteps 4.3.0								

Summary of 60 Measured Person								

	Total		Model		Infit	Outfit		
	Score	Count	Measure	S.E.	MNSQ	ZSTD	MNSQ	ZSTD

MEAN	212.4	5.0	.14	.13	.86	-.32	.87	-.30
SEM	9.3	.0	.17	.00	.10	.16	.10	.16
P.SD	71.2	.0	1.27	.04	.79	1.19	.80	1.19
S.SD	71.8	.0	1.28	.04	.80	1.20	.80	1.20
MAX.	367.0	5.0	3.16	.23	4.30	3.11	4.43	3.18
MIN.	32.0	5.0	-2.89	.08	.03	-2.52	.02	-2.52

Real RMSE	.15	True SD	1.26	Separation	8.21	Person Reliability	.99	
Model RMSE	.14	True SD	1.26	Separation	9.02	Person Reliability	.99	
S.E. of Person Mean	= .17							

Person Raw Score-to-measure correlation = .99								
Cronbach Alpha (KR-20) Person Raw Score "Test" Reliability = .99 SEM = 8.71								

Summary of 5 Measured Item								

	Total		Model		Infit	Outfit		
	Score	Count	Measure	S.E.	MNSQ	ZSTD	MNSQ	ZSTD

MEAN	2548.8	60.0	.00	.04	.91	-.91	.87	-.97
SEM	28.2	.0	.04	.00	.22	1.30	.20	1.10
P.SD	56.4	.0	.07	.00	.45	2.60	.40	2.20
S.SD	63.0	.0	.08	.00	.50	2.91	.45	2.46
MAX.	2617.0	60.0	.13	.04	1.47	2.24	1.42	1.93
MIN.	2445.0	60.0	-.09	.04	.42	-4.00	.42	-3.71

Real RMSE	.04	True SD	.06	Separation	1.55	Item Reliability	.71	
Model RMSE	.04	True SD	.06	Separation	1.73	Item Reliability	.75	
S.E. of item Mean	= .04							

Source: Rasch Programme (Winsteps 4.3.0)

In the case of the example, reliability was high with respect to person reliability (.99). The lower item reliability rate of .75 is still within acceptable range, and was attributed to one marker having a low mean score of 40.75 per script. In practice, this could be adjusted during marker training sessions or by discussion with the moderator or module co-ordinator. This and subsequent iterations have been discussed at greater length in

Chapter 7. The tables and figures above are provided as examples for the purpose of explanation.

5.8.2.3 *Modification of existing scale or development of new scale*

The information gathered from the analysis of the data was used to revise and refine the assessment scale. This was carried out by the panel of five experts (as indicated in Section 5.8.1.3) and included experienced educators and examiners of English at Unisa and other institutions.

The procedure followed was similar to that followed by Hattingh (2009: 173-182), but was adapted to the ODL context, characterised by distance between some of the participants. Thus, communication was conducted by means of face-to-face interaction as well as by means of electronic media such as email and Skype.

McNamara (1996), Saville (2001) and Taylor (2002) agree that developing or reviewing an assessment scale includes three basic stages, namely: the design stage, the construction stage and a trial stage. This procedure was followed for the current research.

5.8.3 The design stage

Prior to the workshop, participants were briefed about the aims of the project and provided with background reading as well as comments from the markers. This preparation assisted them in discussing the efficacy of the rating scale.

During the one-day workshop, the results of the quantitative and qualitative data were discussed in order to determine the type of scale that would be most suitable to assess the construct in the given context. The panel analysed examples of student writing to identify the salient features and distinguish performances at different levels of proficiency. These had been sent to the panel members electronically prior to the workshop, although no marks were provided. At the meeting, marks for the scripts were suggested and reasons given for the marks allocated. This benchmarking exercise was similar to that explained in Section 5.4. The existing scale was then closely examined to

determine whether it met the necessary criteria and how it could be improved, revised – or, as a last resort, replaced. Other scales were also discussed with a view to incorporating features that the panel considered to be relevant. These scales had been used previously by panel members, or were currently being used or suggested by them. These included the multi-trait scale recommended by Hattingh (2009: 276), the grid used for the De Beers English Olympiad, the grid used by the English Department of the Nelson Mandela University, and grids used by the Department of Education. These grids are included in Chapter 7 where the design of the new scales has been discussed. This led to the construction of a first draft of the modified (or new) scale. The procedure commenced at the workshop and was continued during 3 further meetings and via electronic communication such as emails, Skype, cell phone conversations and WhatsApp. These conversations were ongoing until the end of 2018.

5.8.4 The construction stage

The draft scale was then refined and revised, initially at the workshop and then during the course of subsequent meetings and/or communication among the panel members.

5.8.5 Testing of the revised scale

The planning and design process was followed by a trial of the revised scale during which the essays were scored by the original group of markers in order to test reliability, and to evaluate the strengths and weaknesses of the scale. The results of this exercise were quantitatively and qualitatively analysed in the same way as described in Phases 1 and 2 in order to modify and refine the evaluation approach posited by this research. Furthermore, results were discussed at each stage of the process. This provided rich information to reinforce the quantitative aspects of the data.

5.9 ETHICAL CONSIDERATIONS

Permission to conduct the research was obtained from the Ethics Committee of UNISA (Appendices A2), and was based on the considerations in the following sub-sections.

5.9.1 Consent and voluntary participation

The researcher obtained consent from participants (tutors, markers, students and panel members). This was done through the signing of a consent form by these participants (Appendices A2 – A4). Before signing this form, participants received a full and clear explanation of what was expected of them so that they could make informed choices to participate voluntarily (Terre Blanche & Durrheim, 1999: 66). Potential participants were informed that they had the right to discontinue their participation at any point they felt necessary.

Apart from the request for permission to use their assignments (Appendix A2 – A4), an introductory email (Appendix D1) was sent to students, whose scripts had been selected for marking, via the student email service, myLife. The request for permission was attached (Appendices A1 to A4).

5.9.2 No harm to participants

Research should never injure people participating in the study (Babbie & Mouton 2004: 522). For the purpose of this research, voluntary participants were not exposed to any danger to them, their home life, work, friendships, community or any other connections. Research took place in the context of their Unisa studies, in their own homes or places of work and study (i.e. wherever they chose to complete their assignments).

5.9.3 Confidentiality

According to Terre Blanche and Durrheim (1999: 68), students should sign a consent form in which they are assured of the parameters of the confidentiality of any information they supply. These parameters were discussed with them prior to the

research. In this study, no names, addresses or student numbers were used. Each script was allocated a random number, such as “Script 1”.

5.10 OVERVIEW: RELEVANCE OF METHODS TO RESEARCH QUESTIONS

1. As mentioned in Section 5.2 , the study addressed two primary research questions (“Is the existing assessment scale used for the *Foundations in English Literary Studies* (ENG1501) at Unisa valid in terms of the various aspects of validation and purposes [namely formative assessment, summative assessment, feedback]” and “depending on the results of the study, how can the existing scale be modified or replaced in order to produce an empirically validated scale for assessing the essays of the target group?”).

Also, as noted in Section 5.2, these central questions were supported by secondary research questions. The specific methods employed to address each of the secondary questions are discussed below:

Sub-Question 1: What do the results of the empirical research process reveal about the validity of the existing scale?

This question was addressed quantitatively by means of statistical analysis of the results, followed by a qualitative procedure in the form of a discussion of this feedback by the panel. The new (or revised) rating scales were tested by means of a process that had both quantitative and qualitative elements (as discussed under question 5).

Sub-question 2: What are the observations of the tutors and markers who use the scale in order to assess examinations and assessments for this module?

This information was obtained by adopting a qualitative approach and included a study of markers’ comments, as well as the feedback from questionnaires completed by markers and tutors. The questionnaires have been discussed in more detail in Section 5.7 (where the instruments employed in the study have been considered).

The answers were then analysed and the results summarised by the researcher (Section 5.8.1.2).

Sub-question 3: What effect, if any, does the distance learning, multilingual and multicultural context have on the perceived and actual validity of the scale?

This question was addressed mainly in the review and discussion of the theoretical background to research on ODL in Chapter 2. The background to distance learning in which assessment takes place was explored to provide the context of the assessment and validation process. The discussion included research on assessment and on the impact of multilingual, cultural and socio-economic factors on the assessment process in the ODL context.

Furthermore, the unique characteristics of, and problems raised by, the context of the study were borne in mind when considering the results generated by the marking and the subsequent statistical analysis. The question was also addressed in questionnaires sent to panel members, markers of the scripts, Unisa markers and tutors. These are the features that distinguish this study from similar studies that have taken place in a face-to-face environment.

Sub-question 4: What recommendations, principles and insights from other stakeholders can be employed to create an improved scale?

Recommendations from panel members were extracted from the feedback from markers during and after the marking of the scripts, as well as from the results of questionnaires obtained from Unisa markers, tutors and co-ordinators of language modules. The results emanating from this feedback were discussed further by the panel members, and incorporated into the improved scale. Thus, a mainly qualitative approach was adopted, although the statistical analysis was taken into account as a source of evidence on which the recommendations were based.

Sub-question 5: How can the modified or new rating scale be designed and tested to ensure optimum validity?

The new or modified scale was tested by means of a process which followed the same stages as the main procedure. This included:

- marking of a sample of scripts by panel members and markers, using the modified or new scale;
- statistical calibration of the results;
- revising of the scale;
- refining the scale; or
- repeating the process if necessary.

5.11 SUMMARY OF PROCESS

The following is a summary of the validation process followed in this study:

- Review the literature and update the theoretical framework throughout the process.
- Collect scripts.
- Select scripts.
- Obtain consent of participants.
- Copy scripts or save them electronically.
- Brief markers.
- Mark exercise undertaken by the group of markers.
- Analyse data.
- Request and receive feedback from markers and online tutors.
- Obtain feedback from: comments made during interviews, notes, reports and questionnaires, and in the course of discussions at each stage.
- Revise the scale or design a new scale in collaboration with a panel of experts.
- Pilot the new or modified scale.
- Discuss feedback.

- Revise the scale, or
- Refine the scale where necessary.
- Carry out final analysis and interpretation.
- Integrate results and discuss the findings.

The current research used the information obtained from the literature reviewed, as well as that extracted from the empirical data to validate an assessment scale that meets the needs of the target group. Although the process seems linear, as depicted above, its implementation was recursive, as indicated by the timeline in Section 5.3.

During the research process, the challenge of reconciling the criteria stated in the relevant outcomes (see Appendix C) with the characteristics and needs of the diverse target group was considered (Chapter 2). This led to a clarification of the outcomes, and a matching process to align the recommended scale(s) to the stated criteria (Chapter 6).

5.12 CONCLUSION

In this chapter, an account is given of the research design and method followed in this study to validate and develop or modify a rating scale for assessing the essays of students enrolled for the module on *Foundations in English Literary Studies* ENG1501. It was emphasised that an empirical approach was central to this validation process, and that a combination of qualitative and quantitative elements was employed to provide sufficient evidence and sources of rich data. After discussing the sampling method, each of the phases was briefly described. Ethical considerations were then discussed, followed by an overview of the research questions and a final brief summary of the process of the research project.

CHAPTER 6: RESEARCH FINDINGS: EXISTING SCALE

6.1 INTRODUCTION

The findings of the research carried out to determine the validity of the existing rating scale are discussed in this chapter. The chapter begins with the results of the pilot study, which include the quantitative findings and qualitative findings in the form of comments made by markers during this phase of the process. After this, the results of the main study are discussed in detail. The results of the marking were analysed quantitatively, using statistical calculations to measure the validity of the scale, including scoring validity or reliability. The findings of the qualitative research are then discussed. These include comments from the markers involved in this study as well as from Unisa online tutors and markers of ENG1501. Issues arising from the process have been discussed after each stage and the findings are then summarised at the end of the chapter. This prepared the way for the next phase, namely the development and trial of alternative scales.

6.2 SELECTION AND DOWNLOADING OF SCRIPTS

As described in Chapter 5, the student population investigated in this study comprised the students who were registered for the *Foundations in English Literary Studies* (ENG1501) module at the University of South Africa and submitted the first assignment that was set for the second semester of 2016 as part of their formative assessment. The target assignment was based on a prescribed poem and thus tested the students' knowledge of and insight into this particular genre.

6.3 PILOT STUDY

The purpose of the pilot study was to test the planned procedure and instruments, by marking a small sample ($n = 20$) of scripts, followed by a quantitative and qualitative

analysis of the results. The procedure that was followed has been described in Section 5.8.

6.3.1 Statistical results – analysis of variance (20 scripts)

After the marking had been completed, the data were statistically processed to test scoring validity and the extent of inter-rater reliability. The three features analysed were content, language use and the total mark. Unfortunately, results for Marker 1, the original Unisa marker, were available only for the total mark. This applied to the main study as well. As pointed out in Chapter 5 (Section 5.8), the reason for this marker awarding just a single mark would seem to be the time constraints affecting the delivery of this module, exacerbated by the logistical and administrative problems generated by the distance learning environment.

For these reasons, the participating Unisa markers gave only a final mark, although the grid was provided to them and they were required to consult it, at least as a guideline. Thus, the pilot data for the total mark included all five markers, including the Unisa marker, but the data for content and language use reflected the marks of the other four markers only.

As regards the total mark, the reliability of the markers (an aspect of validity as defined in Chapter 3) was determined by means of SPSS, with the Cronbach Alpha of 0.962 and the intra-class correlation of 0.862 (assuming a two-way ANOVA-model with the scripts random and markers fixed) and 95% confidence limits of 0.753 - 0.936. As explained in Chapter 5 (Section 5.8), the Cronbach Alpha reliability index is scaled from 0 to 1, with values closer to 1 indicating good reliability. These results were largely positive, particularly the Cronbach Alpha result which indicates high consistency among the markers (0.962). The intra-class correlation of 0.862 is also considered to be good to excellent (Cicchetti, 1994). The confidence limits of 0.753 – 0.936 indicate the upper and lower limit for the mean score (the narrower the interval, the more precise is the estimate). This was relatively wide and indicated a wide range for the intra-class correlation. This was further investigated using the vertical maps (showing the ability range of students) employed in the main study.

6.3.2 Comments from pilot study markers

The group of markers was requested to submit comments on the marking process. These comments are now presented verbatim. Issues that emerged at this early stage of the process were the following.

6.3.2.1 *The range within some of the levels, particularly Level 3.*

It was noted that all markers commented on the need for more levels, particularly regarding Level 3, with its large range (32% - 54%), which created difficulties in distinguishing between a pass and a failure mark. All markers felt that Level 3 should be divided into two levels to indicate the difference between “pass” (50% and above) and “failure” (below 50%). This is a serious weakness of the existing scale, as the problematic range could result in markers having to judge a pass or failure mark without any guidance from the criteria on the marking grid.

Marker’s comments:

- “Having only four levels is proving to be a problem as the range within the level is problematic”.
- Comment on a script: “The student gets halfway in understanding, but there is no category like that anywhere so one has to twist it into another category. Not sure that matches the 50% awarded to others who have received 50%, but I was going by the categories of the grid”.

6.3.2.2 *Lack of sufficient descriptors*

Markers commented on the lack of descriptors. This issue could be linked to the number of levels, since additional levels would give rise to more (and possibly more precise) descriptors.

Markers’ comments:

- Comment on a script (language mark): “I want to put this at level 4, but none of the descriptors allows me to do so. Consequently, this mark is inflated”.

- “I am not marking as I would with a set of papers, where I would constantly return to the previous papers to see if I were happy with the mark for the present piece in relation to those previously assessed. If I did this, I would have to fiddle with the marking grid. This is perhaps a comment worth making on the use of grids. Although I might be able to abide by its strictures, the grid would have to provide me with more descriptors so that I could make such fine distinctions”.

6.3.2.3 *Difficulty in distinguishing between content and language use*

The relationship between content and language use appeared to be far more complex than was reflected by the simple allocation of equal total marks to each of these components. This relationship was discovered to be considerably more interlinked than the current marking grid indicated.

Markers’ comments:

- Comment on a script: “Difficult to distinguish between content and language as the pomposity of the style requires the student to make sweeping statements and to wander into over-interpretation. There is something sound in this response, but it is buried so deep one tends not to see it. This is more polemic than answer!”
- Comment on a script: “This seems to be the worst assignment, perhaps because the [poor] language [usage] makes the content seem trite.”

These were the issues that signalled initial reservations about the existing marking grid and were explored at greater length once the results of the main study were available. At this stage, the basis for the panel discussions was laid following the marking of the sample scripts.

6.4 MAIN STUDY (60 SCRIPTS, INCLUDING SCRIPTS 1-20)

Once the pilot study had been concluded, the main study commenced, and progressed as follows:

6.4.1 Collect data and mark scripts

As described in Section 5.8, randomly numbered copies of the selected scripts, from which the original marks, students' names and student numbers had been removed, were sent to markers who marked them, using the existing rating scale (Section 5.8).

6.4.2 Analyse data

After the marking process, the data were analysed statistically, using the FACETS version of the multi-faceted Rasch programme (Linacre 2006b) and Cronbach Alpha, as for the pilot test.

6.4.3 Explanation of the results

The explanation of the results is based on that given by Van der Walt and Steyn (2007) and Hattingh (2009: 171-173). The detailed tabulations for content and language use have been provided in Appendix F, but the results have been summarised below.

A summary of the estimates (including a reliability index) are found at the bottom of each of the following tables. The reliability index reflects the extent to which the scale defines different ability levels, and distinguishes between the assessed performances (McNamara, 1996: 138). Differences between raters were to be expected and, despite training, it is not possible (or maybe even desirable) to eliminate all differences in rater scoring (McNamara 1996: 140). Another consideration is that, in practice, during marker training sessions, markers who are identified as marking too high or too low are trained until they are able to adjust their scores.

6.4.3.1 Markers 1 - 6 (scripts 1 - 60)

The following information was generated by the Rasch analysis for Markers 1 - 6 (Totals) and Markers 2 - 6 (Content and Language). This was the most complete data set. The information is summarised in the following tables, and a more detailed analysis is provided in Appendix F. Total scores included those given by the Unisa marker (Marker 1). Figures 6.1 and 6.2 below present the distribution of raters (items), as well as persons (test taker scores). As explained in Section 5.8.2.2, the maps indicate the degree of inter-rater consistency as well as their degree of scoring accuracy and severity of marking. Test-takers placed above the 0 logit mark display higher ability, while those below the 0 mark display lower ability.

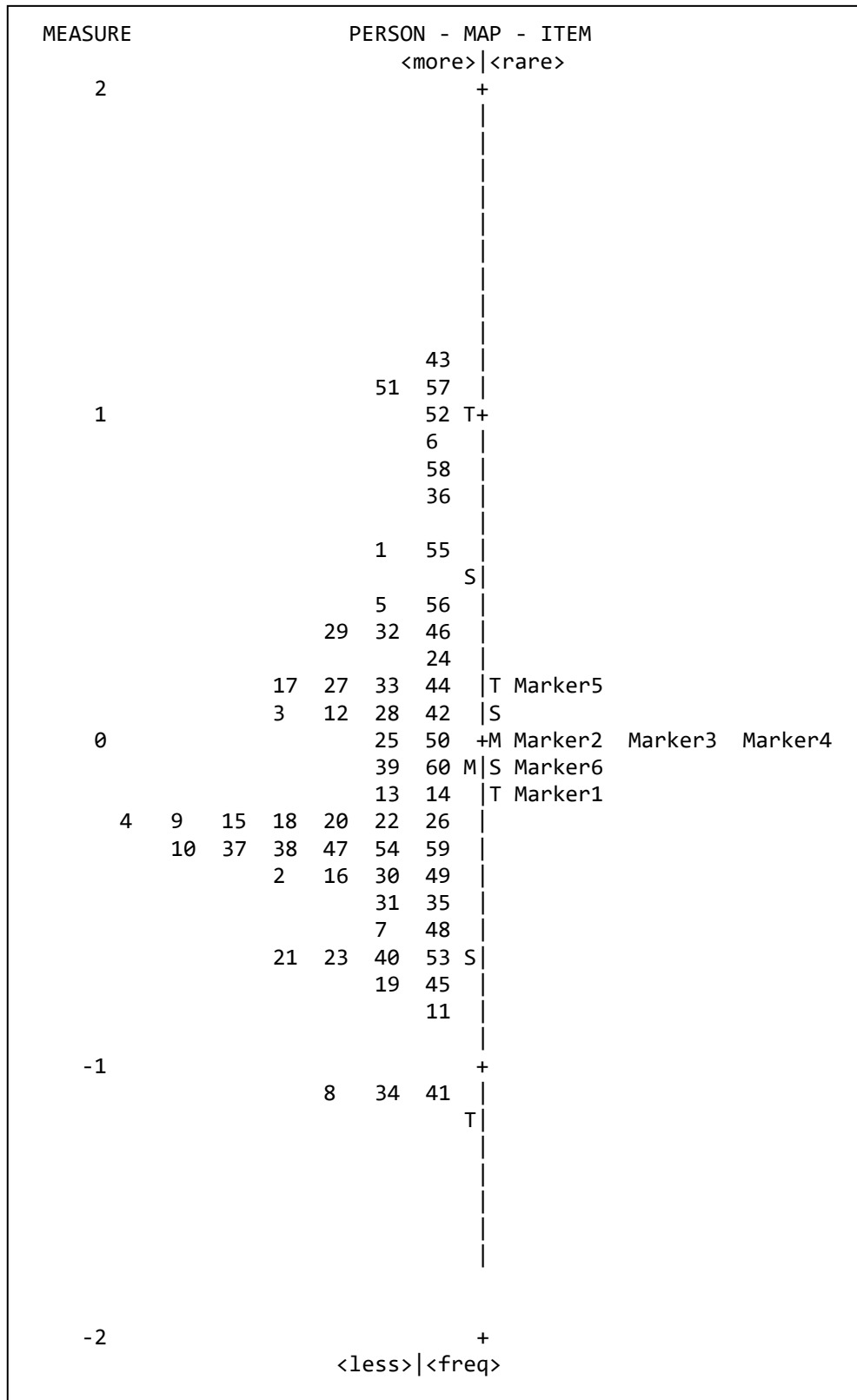


Figure 6.1: Distribution of items and persons: existing scale, Markers 1 – 6 (60 scripts)

In Figure 6.1, most markers were clustered around the logit 0 mark. Marker 5 was higher (more severe) than the others, but still within satisfactory range. The same

applied to Markers 6 and 1, who were slightly below the rest, but did not reach the -1 mark. Thus, the consistency among markers could be rated as good. The scripts ('persons') ranged between 1 and -1, although most clustered just below the 0 logit mark. This indicates a fairly wide ability range.

There was some concern that data for Marker 1 was available only for the total mark. Figure 6.2 below shows the distribution without Marker 1.

In Figure 6.2, the distribution was not affected by the omission of Marker 1, who had been plotted, along with Marker 6, as slightly lower than the rest, but still close to the 0 logit. The consistency between raters remained good and the test-takers (persons) continued to be ranged between 1 and -1, although most clustered just below the 0 logit mark. Thus, the consistency among markers could be rated as good, and the scripts ('persons') range was similar to the ability range found in Figure 6.1.

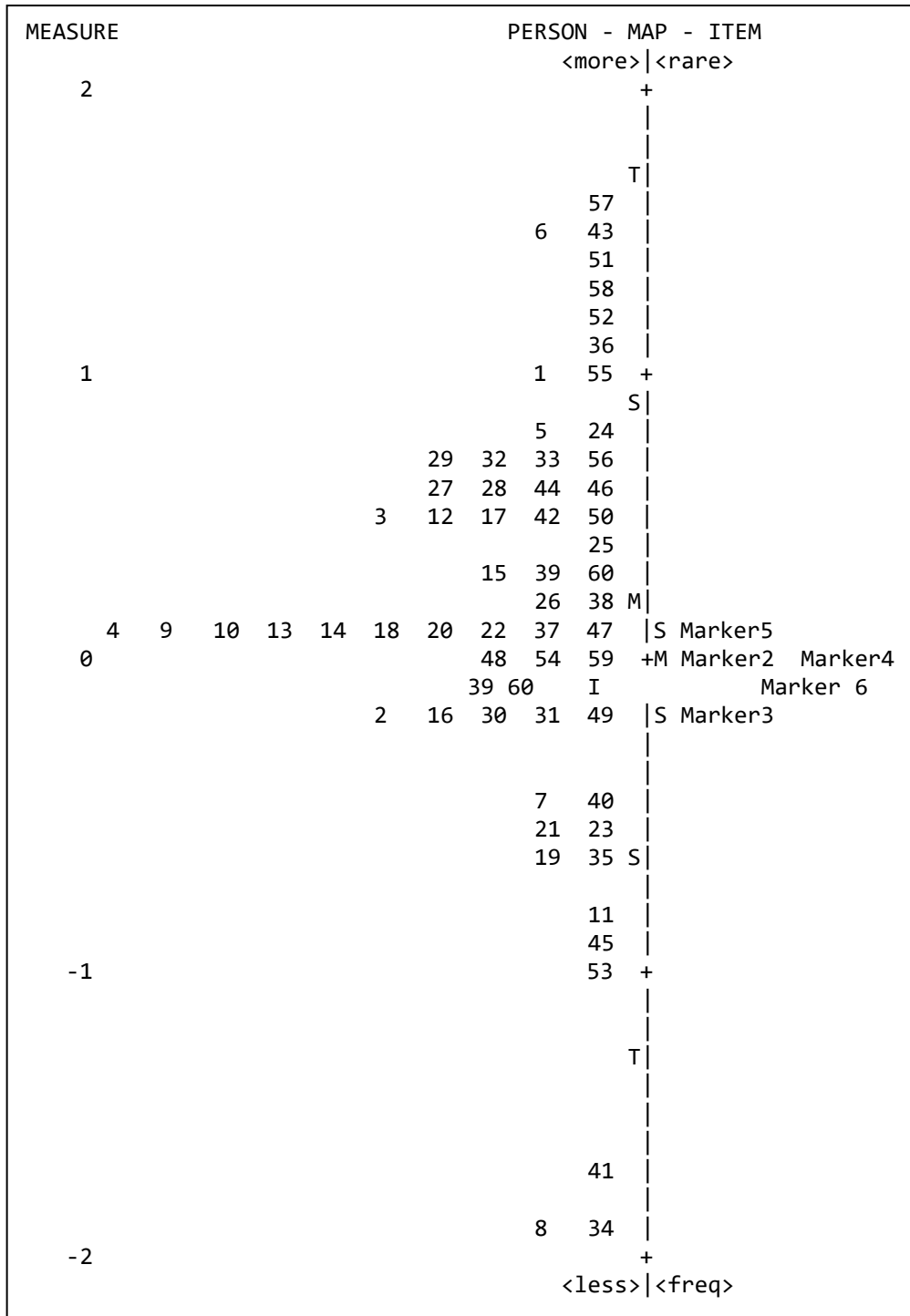


Figure 6.2: Distribution of items and persons: existing scale, Markers 2 - 6 (60 scripts)

In the case of Markers 1 – 8 presented below (Figure 6.3), the “Person” column is slightly closer to the 0 logit mark, while the “Item” (or Marker column) remains clustered around the 0 logit mark, once again indicating good marker agreement and consistency. The omission of Marker 1 did not affect this consistency (Figure 6.4).

The following figure (Figure 6.3) shows the distribution of items and persons: existing scale, Markers 1 - 8 (30 scripts)

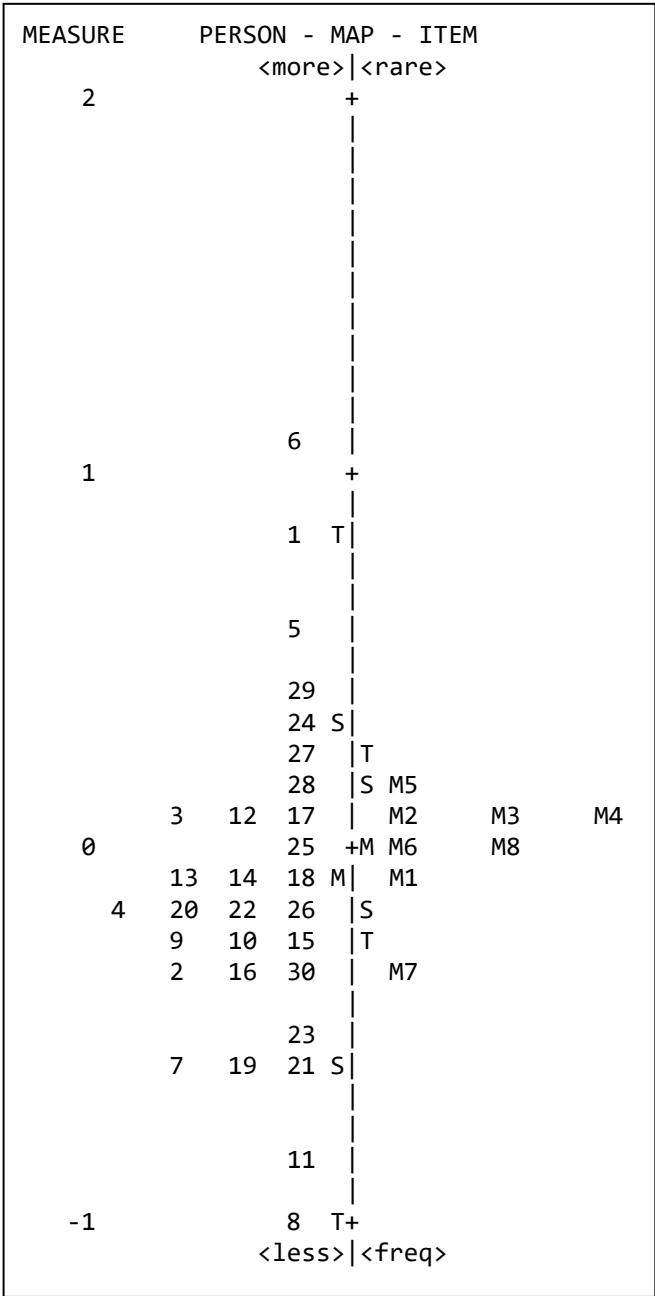


Figure 6.3: Distribution of items and persons: existing scale, Markers 1 – 8 (30 scripts)

In Figure 6.4 below, Marker 1 has once again been omitted. The figure thus shows the distribution of items and persons for Markers 2 – 8 (30 scripts).

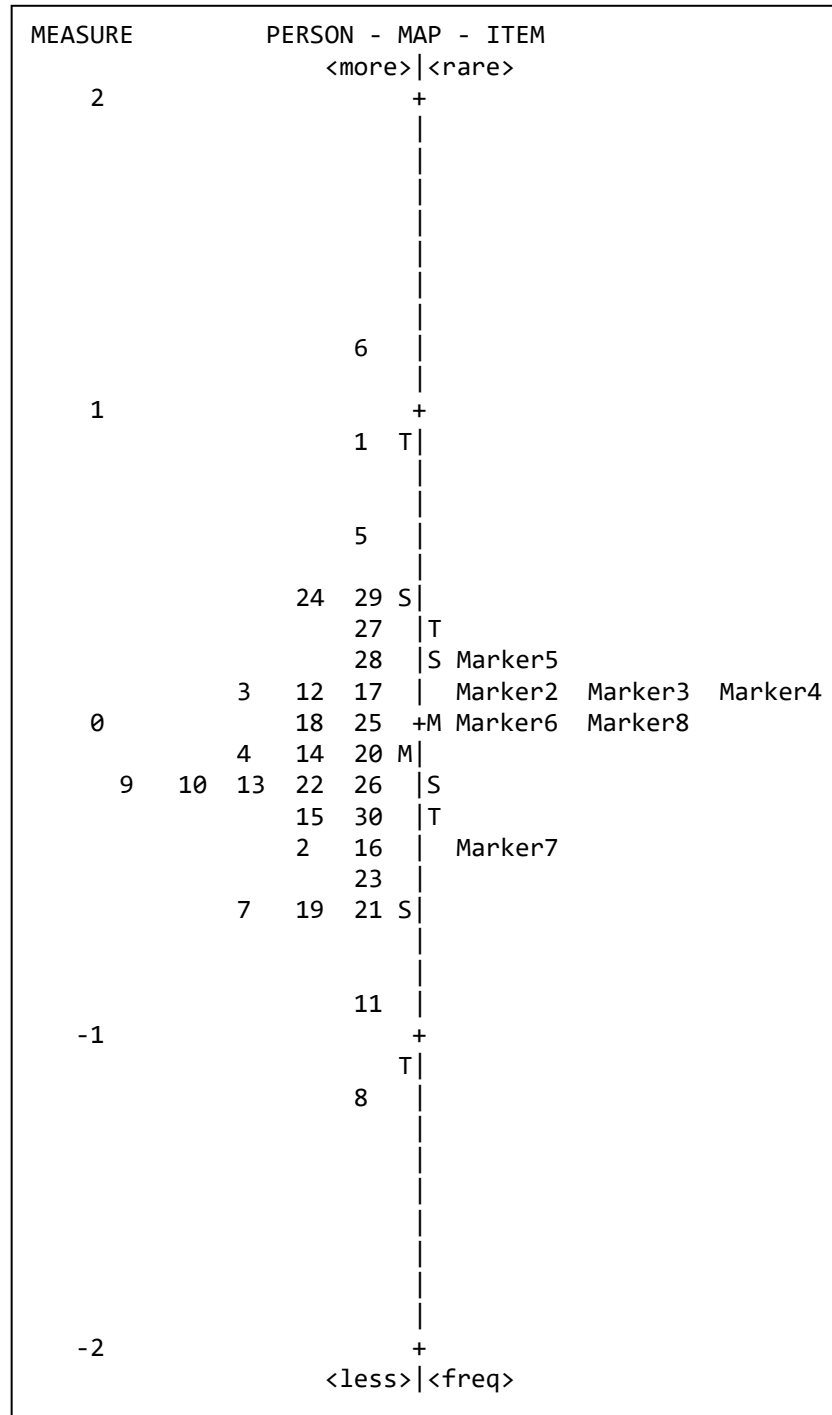


Figure 6.4: Distribution of items and persons: existing scale, Markers 2 – 8 (30 scripts)

In the next figure (Figure 6.5), the last 30 scripts (Scripts 30-60), marked by Markers 1-6, 9, and 10 were statistically analysed. Again, the team of Markers 1 to 6 was retained, with the addition of Markers 9 and 10, who marked scripts 31 to 60. Once more, the clustering of markers was satisfactorily around the 0 logit. A cluster was apparent in the “Person” column, with all scripts in a loose cluster around the 0 to +1 level.

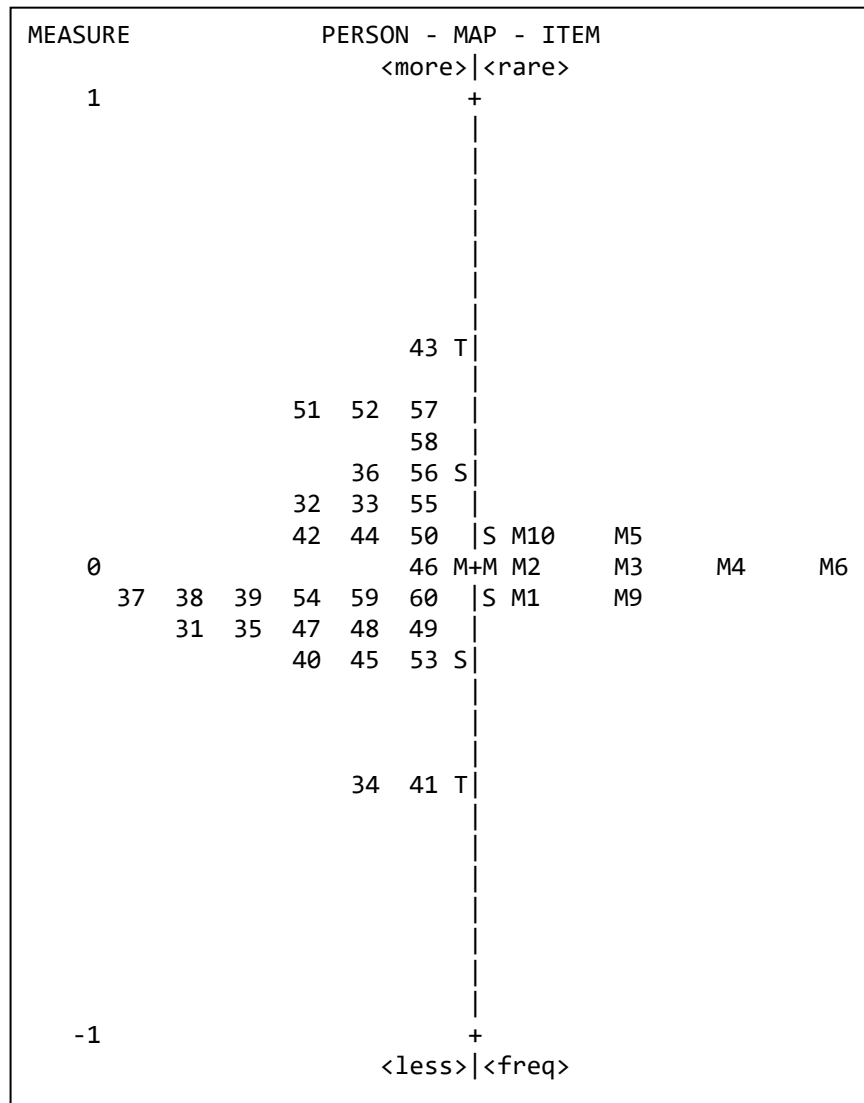


Figure 6.5: Distribution of items and persons: existing scale, Markers 1, 6, 9, 10 (30 scripts)

6.4.3.2 Total marks: script and marker reliability

The total marks of Markers 1-10 were then investigated for ‘person’ (i.e. script) and ‘item’ (i.e. marker) reliability. Scores close to 1.00 indicate high reliability, and scores from 0.75 are considered good to acceptable. The results have been summarised in Table 6.1 and details have been provided in Appendix F.

Table 6.1: Script and marker reliability: Total marks, Markers 1 – 10 (60 scripts)

Script numbers	Marker	Script Reliability	Cronbach Alpha (Raw score-to-measure correlation)	“Test” Reliability	Marker reliability
Scripts 1- 60	1- 6	.91	.99	.92	.95
	2- 6	.97	.98	.96	.81
Scripts 1-30	1- 8	.97	.99	.97	.91
	2- 8	.97	1.00	.96	.94
Scripts 31- 60	1- 6, 9,10	.95	1.00	.95	.70
	2- 6, 9	.97	1.00	.96	.79

The Cronbach Alpha scores were high to very high (i.e. 1.00 to 0.98) and the scores in general indicated excellent reliability, with the exception of the marker reliability of Scripts 31 - 60 (Markers 1 - 6, 9, 10). Without Marker 1, the reliability for markers in this group was low (0.65) and it seemed that this was due to Marker 10 whose results produced a large infit-value of 2.79. The analysis was thus repeated without this marker, resulting in a higher reliability (0.79). Apart from this batch of scripts, it was felt that the sample generally showed good reliability levels, that marker consistency was acceptable, and that the scale gave a fair reflection of students’ abilities since the script reliability was high (i.e. 0.91, 0.97, and 0.95 for the respective batches of scripts).

6.4.3.3 Content/organisation and language use/style

An analysis of the content and language use scores was then conducted. This was done by taking the scores for content and language use of Markers 2 - 6 for both Language and Content as 10 “ITEMS”, labelled M1_L, M1_C, etc. The relevant tables are provided in Appendix F.

In summary, the results of the exercise were generally positive. The sample was considered sufficiently representative of examples of the various levels and also indicated the validity of the scale. The reliability indices were positive in most cases (between 0.92 and 0.94).

This statistical analysis was supplemented by qualitative evidence in the form of comments by markers, and questionnaires distributed to tutors and markers at Unisa (Section 6.4.5). The markers of the main study of this research were also asked for

comments at the conclusion of the marking process (Section 6.4.4). These are discussed in the following section.

6.4.4 Comments by markers of the main study

After marking had been completed, markers were asked to submit comments about the marking scale. These gave rise to the following observations.

6.4.4.1 *Number and range of levels*

The concern about the number and range of levels on the assessment scale was repeated by several markers. The wide range represented by Level 3 posed a problem in distinguishing between a pass and a failure mark. Markers strongly recommended that Level 3 be split into two levels, namely one level to indicate 50% and above (i.e. a pass mark) and a lower level below 50% to indicate “failure”.

Marker’s comment:

- “If only we can find a way of giving ourselves a clear distinction between those ... candidates that fall a fraction above or below the pass mark”.

6.4.4.2 *The weighting of marks for content and language use*

Giving equal weighting to content and language use was seen as problematic, and gave rise to questions about the relative importance of the two outcomes for the module (Appendix C). The relationship between the content and language use was seen to be less simplistic than the marking grid seemed to reflect. A concern affecting the validity of the scale was that, when the marks were added together for a final result, it was theoretically possible to pass the assignment with a good mark for language use, while demonstrating an inadequate understanding of the poem. It was agreed that the relative weighting of the scale would need to be addressed.

Marker's comment:

- “Another observation about the validity of using the grid as the measuring instrument... is whether it measures what the assignment was intended to evaluate. Theoretically, somebody with poor understanding of the poem who expresses it very well could score better than somebody who expresses an understanding of the poem very badly. Is the assignment evaluating use of English primarily or primarily understanding and literary appreciation of the poem?”

6.4.4.3 Plagiarism and referencing

It was noted during discussions with panel members and markers of the partial pilot study that the existing marking grid makes no reference to plagiarism and, thus gives no indication of how this should be addressed. This topic was also raised during the main study where cases of plagiarism were encountered. Panel members and markers were divided with regard to the severity of the penalty, but all agreed that the rating scale should indicate that some penalty be incurred, especially since plagiarism is widespread in tertiary education and particularly in the ODL context (Tennant *et al.*, 2010: 94; Boughey, 2013: 31 – 33, Minnaar, 2012). For instance, Minnaar (2012: 6) states that:

...the ODL institutions themselves are making dishonesty easier through their increasing use of technology in teaching and learning. New electronic tools and technologies are making it convenient for students to be dishonest and plagiarise other people's work.

Other factors cited by Minnaar (2012: 5) are:

Fear of failing the examination, inadequate preparation, lack of confidence, knowledge of dishonesty by other students, the need to earn a good pass mark, the desire to avoid being disgraced in front of family members, forgetfulness and the urge for promotion in the workplace.

These factors are prevalent in ODL where students are studying in a non-institutional environment (such as the home or workplace), the stakes are high and there is very little FTF guidance from lecturers, tutors or even peers, although the study material does caution against it.

Markers of the main study cited plagiarism as a problem, this time in conjunction with referencing (or the lack thereof). The question was raised as to whether defined penalties should be imposed for incorrect referencing in a first assignment (Boughey 2013: 31), but markers and panel members were unanimous that plagiarism must be addressed by the scale. An apposite remark that sums up the issue was a marker's opinion that, giving credit to "the sometimes weak attempts of a student who has really tried to engage with the text", is preferable to "giving marks to a student who has settled for copying random sections of another scholar's work, without attempting to understand it and/or to come to some conclusion for him/herself". Another comment on a script further illustrates the relationship between content and style with respect to plagiarism:

- "[The student] has swallowed the feminist dictionary but is unable to relate it to the poem at hand so makes sweeping statements. [This is][p]ure ideology, not a sensitive reading of a poem. Not sure if this is style copying, though much of the content is plagiarised. S/he has at least attempted to relate this to [the] background material. If s/he weren't so full of polemic s/he might have achieved a good mark".

Another marker commented:

- "I feel... that the rubric is lacking in that it does not take the issue of plagiarism into account. There is clear evidence in many of these scripts that the students have either lifted the work of others directly, or have used quotation marks, but with no referencing (there was not a single script where any inline referencing occurred). How then is a marker supposed to credit the sometimes weak attempts of a student who has really tried to engage with the text? In my mind that is far worthier than giving marks to a student who has settled for copying

random sections of another scholar's work, without attempting to understand it and/or to come to some conclusion for him/herself?”

6.4.5 Questionnaires

The purpose of the questionnaires sent to tutors and markers of ENG1501 at Unisa was to find answers to the research questions posed by this study (Chapters 1 and 5). The following research questions were of particular importance at this stage:

- What are the observations of the tutors and markers who use the existing scale to assess examinations and assignments for this module?
- What effect, if any, does the distance learning, multilingual and multicultural context have on the perceived and actual validity of the scale?
- What recommendations, principles and insights from other stakeholders can be employed to create an improved scale?

In designing the questionnaires, it was stressed that practical suggestions should be offered and that the participants should be encouraged to express their opinions and experiences of the scale. Thus, the emphasis fell on open-ended questions and, although closed-ended questions were also asked, each of these provided space for comments. The responses provided a source of rich and valuable information, and many of these comments were followed up by emails from and to the researcher (Examples in Appendix G).

Tutors and markers were asked the same questions, with the exception of Question 6 for markers which pertained to their approach to the marking of scripts. In total, 11 tutors submitted answers, while feedback was obtained from 7 markers. The following summary is based on the responses received.

6.4.6 Summary of feedback

Table 6.2 shows a summary of the results and the agreement between the two groups (i.e. tutors and markers). This summary is followed by the researcher's observations (Section 6.5)

Table 6.2: Summary of responses to questionnaire (existing scale)

<p>1. Question Do you think that the scale adequately assesses the construct of the module as stated in the outcomes (i.e. does it test what it is supposed to test)?</p>
<p>Results</p> <p>The majority (72.2%) agreed with the statement. Tutors: Yes: 7 (63.6%) Markers: Yes: 6 (85.7%)</p>
<p>Representative comments</p> <p>Yes</p> <ol style="list-style-type: none"> "The grid tests what is required by the module outcomes and what is required from students to have learned. For me there are three elements that are essential for literary analysis: language, structure, and analysis, which the scale assesses". "This is a foundation module so students are introduced to a variety of texts to get an overall view. The scale assesses this adequately". "The scale assesses the three essential aspects of the modules construct, namely, lexicon, semiotics and the functional grammar of the English Language". "I think the scale assesses all the important aspects of ENG1501, as it addresses the content /organisation and form in which the content is presented. It is important in a literature module to address both <i>what</i> is said, and <i>how</i> it is being said". "The current scale, in my opinion, covers the module outcomes well, and does so in a manner that is fair". "It gives equal weight to the desired outcome (constructing a well-structured argument) and applicable language skills". <p>Partially</p> <ol style="list-style-type: none"> "I have a huge problem with the term "shaky" in the last two levels. What does 'shaky' even mean? It is certainly too colloquial a term to include in a formal, academic assessment scale. What are the grounds for calling something 'shaky'?" "The scale also allows for a huge amount of subjective interpretation on the side of the marker... Having said that, I do realise that subjective interpretation on the side of the marker will always play a role in the marking of English Literature where there are often no correct and incorrect answers, only well-motivated and poorly motivated ones". "Sometimes lacks specificity". "Level 3 for content/organisation and for form is too broad. There is a huge gap between 54% and 32%".

<p>2. Question</p> <p>In your opinion, is the distinction between the band levels clear?</p>
<p>Results</p> <p>The majority (61.1%) agreed. Tutors: 6 (54.5%) Markers: 5 (71.4%)</p>
<p>Representative comments:</p> <p>Respondents felt that the distinction was clear, although there were problems with the levels.</p> <p>Yes</p> <ol style="list-style-type: none"> Clear distinction of each band and the requirements for that band. Criteria explanation clear”. “There seems to be a clear distinction between the levels, with level 1 (for example) representing a distinction i.e. an excellent grasp of the material as well as the ability to communicate this understanding”. “The score divisions, e.g. 100% -76% etc. are divided fairly and in such a way that provision is made for every type of student, no matter their skillset. It also makes it easier for students of different skill levels to be assessed fairly”. “The four categories relate closely to the way in which one critically assesses work naturally, and each has a number of factors which guide where the mark should be within it”. <p>No</p> <ol style="list-style-type: none"> It will always be very difficult to have clear distinction between the levels ... when it comes to many subjects in the humanities field. What are the criteria for measuring vocabulary as that of a ‘sophisticated range’ for instance? This is solely a subjective interpretation”. “The band levels could use further explanation to act as a guideline for both student as well as assessor. An example of this is specifically focused on the borderline cases where critique and informed commentary is necessary for students to improve on their writing. Where a case receives between 54% - 32%, I often provide slightly more in-depth commentary to my face-to-face students [at another university]”. “Particularly for the lower mark allocations, it becomes important for both tutor and student to understand why a mark allocation would fall in a particular category”. “Level 3 is confusing because the student can pass or fail on this level but the content and organisation does not change”. <p>Partially</p> <ol style="list-style-type: none"> “‘Vocabulary’ and ‘language usage’ do not seem to be very distinct categories because vocabulary is inherently part of language usage therefore the distinction seems superfluous. For example, ‘mastery of word form’ (vocabulary) and ‘effective complex constructions’ (language usage) could become a single mode of assessment thereby making the categories more succinct. I concede, however, that vocabulary is different from language usage in terms of sentence structure and overall coherence”.
<p>3. Question</p> <p>Are there sufficient levels? Choose one of the following answers.</p> <p>Yes No, too many No, too few</p>

Results

The majority (66.6%) believed that there were ‘too few’ levels. The breakdown was as follows:

Tutors: 6 (54.5%)

Markers: 6 (85.7%)

6 respondents (38.8% i.e. 5 tutors and 1 marker) felt that there were “sufficient” levels, while none were of the opinion that there were “too many”.

Representative comments

Yes

- a. “The levels give a good overview of range. However, I think that 32% is too low to be considered ‘shaky’ and this is a vague term as well. 32% is a strong fail and should be considered ‘very shaky’”.
- b. “I am not convinced that markers actually use the grid at all. Examiners are likely to give marks based on an overall impression of the work. A good examiner will likely know if the student deserves a good mark or not based on the student’s ability to formulate a proper argument, their usage of proper vocabulary, grammar and spelling. Examiners will probably not check their impression against the scale in a ‘real life’ situation. (I am not an examiner and I am not involved with the marking of any Unisa scripts. I say this on the basis of my own personal experience at another institution.)”.
- c. “Adding more levels will make it hard to work efficiently through large numbers of assignments (for markers/moderators). The current scale still offers flexibility because of the range of marks contained in each level of the scale”.
- d. “If there were more levels, it might make it more difficult for e.g. markers to assess which category a student needs to be placed in, as the decision scope would be too broad. The “at risk” level could also be helpful in identifying which students need more help/attention”.

No, too few (majority judgement)

- a. “I am in favour of a separate level which includes referencing and style apart”.
- b. “I feel there should be another level between 75% and 100% to encourage students to aim higher (and perhaps markers as well)”.
- c. “I think there needs to be a 5th band so that mark allocation is not so broad. For instance, Level 2 is ‘good to average’ but there is a marked difference between 56% and 75%. There should be allocation for 60% - 69%”.
- d. “From personal experience, student’s abilities with writing can be very complex. This invariably means that a more concise guideline with far more consideration towards writing styles and subjective interpretations of texts need to be catered for. What I am trying to say is that mark allocation happens in consideration of a number of factors that are at play at once. It would be far more helpful to have more levels to cater for students who understand the work thoroughly but have not yet developed a language to effectively describe their thoughts within their analyses”.
- e. “I would prefer 5 levels, with level 3 being divided into 2. This is because I think that grouping 54%-32% together is a bit too broad, especially since this groups those failing an aspect of the module with those who are just passing it”.
- b. “Level 3 needs to be broken into two levels. A student cannot pass or fail based on the same level, it does not make sense. It’s like playing roll the dice”.

4. Question

Do you agree with the present 50/50 weighting of marks between organisation/content and language use?

Results

Most tutors were in agreement, but the markers were undecided. The combined total in agreement was 55.5%

Tutors: 7 (63.6%)
Markers: 3 (42.8%)

Representative comments

Yes

- a. "The study of English Literature at Unisa does not happen in isolation.... If there is a universal expectation that students of English Literature should be able to structure their answers logically and well, the same expectation should be upheld at Unisa".
- b. "The module assesses both content (literary studies) and language (writing and language competency). Both are equally important in this course".
- c. "I agree with the given weighting of marks as it is a simpler way of providing an outline for first year students who are still learning about literary criticism and academic writing. Often, they do not come to university equipped with the organisational skills to construct academic papers. This suggested weighting simplifies this complex issue with students who are still in the learning process".
- d. "A large part of student writing and the clarity/understandability thereof does rely on language, which is why I think the 50/50 weighting is fair. If a student's language use is faulty, it can often obscure the meaning of their content and will influence the organisation as well".
- e. The 50/50 weighting is fine; it is the levels that are confusing".

No

- a. "A third level which distinguishes referencing separately should be set apart. Referencing should not be grouped with all other so-called mechanics".
- b. "No. This is a tough one, because it is an English module and therefore seems to preclude good English writing skills, but so often this is not the type of student we encounter. The fact that this is also distance education, I think, could maybe be a factor in the seemingly low quality of written English we encounter in students. As a result, I think there should be a 60/40 weighting, which isn't much, but which might be beneficial in the long run".
- c. "I think that a 33/33/33 weighting of marks between content, structure, and language is better, as it makes it somewhat less ambiguous for both markers and students. (not that this is not addressed to a degree in the bands themselves)".

5. Question

Are there any features of the scale that you think are open to misinterpretation or subjectivity?

Results

Undecided
50% disagreed.
Tutors:
Yes: 5 (45.4%)
No: 6 (54.5%)
Markers:
Yes: 4 (36.3%)
No: 3 (42.8%)

Representative comments

Yes

- a. "It seems strange to give two separate marks for an essay instead of looking at how all the criteria work together to make a good argumentative essay".
- b. "Yes, pretty much all the criteria are dependent on a subjective interpretation. No two examiners will give the same paper/student exactly the same marks. This is not a unique circumstance due to the given scale used at Unisa. Examiners of English Literature all over the world will agree that there is a subjective element to the marking of questions in this

<p>field of study”.</p> <p>c. “Distinction and above. An extra level for the top students (75% - 100%)”.</p> <p>d. “As previously stated, the lower rankings of the scales need more concise reasoning in order to act as a guide for both student and tutor about the expected criteria and quality of work expected at a tertiary institution. The less there is for students and tutors to go by, the more the scales are open to subjective interpretations”.</p> <p>e. “Range of vocabulary (this is subjective and largely irrelevant). Mechanics and organisation seem largely interchangeable”.</p> <p>f. “Level 3 is open to misinterpretation and subjectivity. The marker can either pass or fail the student. This category is too broad”.</p> <p>No</p> <p>a. “Not if markers/moderators are adequately qualified and trained”.</p>

<p>6. Question</p> <p>What is your preferred approach when you mark ENG1501 assignments/examination answers? (for markers only)</p>
<p>Results</p> <p>I use the scale as a guideline, but use my own discretion: 1(14%) It depends on “circumstances”: 5 (71%)</p>
<p>Representative comments</p> <p>I use the scale as a guideline</p> <p>a. “By referring to the scale from time to time, I am able to avoid becoming too harsh or lenient. I am able to check myself in terms of where a language mark should be awarded”.</p> <p>It depends on circumstances</p> <p>a. “For the majority of assignments, I adhere strictly to the rating scale. I make an exception only if an assignment is exceptional in either content or organisation. For example, if the arguments are exceedingly well-developed and show critical and original thought, I might award a mark on the lower end of Level 1 even if the organisation lacks some features. Likewise, if the organisation is impeccable I might award 38 or 39 out of 50 even if the arguments could have been more thoroughly developed. Similarly, flawless language use might be awarded one or two marks more than the content strictly warrants. But these exceptions are very few and far between, perhaps two or three out of a hundred”.</p> <p>b. “I use the combined marking grid which I created [based on the existing scale], because I find it easier to read my version at a glance”.</p>

<p>7. (Question 6 for tutors)</p> <p>Should the scale be designed to take the multicultural and multilinguistic distance learning target market into account?</p>
<p>Results</p> <p>Most agreed Tutors: 7 (63.6%) Markers: 5 (71.4%)</p>
<p>Representative comments</p> <p>Yes</p>

- a. "This is something that I think about often. At university, especially for a module on English Literature, we tend to assess the students' work through a western gaze that is designed to enforce a western standard consequently excluding many. This is, however, the contradiction of this module. What I think is useful to take into account is the student's ability to analyse a text based on how they are instructed to do so. As any other module, there is content for the student to learn. In this module, we teach them how to analyse in a particular way and therefore, ideally, they should be able to do so despite diverse backgrounds. A weight of 25 language, 25 structure, and 50 analysis is perhaps a good way to assess their work".
- b. "Home language [influence] should be allowed where it does not obscure meaning completely".
- c. "The fact that ENG1501 is a distance-learning module necessitates the accounting-for of the multicultural and multilinguistic distance learning target market. We have a pretty good idea who our students are, and I think there are steps that can be taken to be more accommodating. I think it is very important to uphold the integrity of the module, but if students are going to be almost-arbitrarily passed in order to meet stipulated pass rates, then perhaps something like the scale should be adjusted so that passing is less arbitrary".
- d. "It is quite simple: multicultural, multilinguistic and distance learning target market is the target market of an institution such as UNISA. If you do not cater to the context of situation and culture of your students, then we will be failing our students as educators and an educational institution".
- e. "I think that one should be sensitive to these issues in a South African context, but that one should also be cautious about any kind of reduction in scope or difficulty. It has to be carefully considered, as the students should be internationally competitive".
- f. "While I think the scale does take these factors into account on some level already, it is important to consider the multicultural and multilinguistic distance learning target market if one keeps in mind the very nature of UNISA's purpose and its learning system, as well as the broader market it services".
- g. "The scale should be designed to take the multicultural and multilinguistic distance learning target market into account for at least the 1st assignment. Give the students detailed guidelines on how to improve on the 2nd assignment".

No

- a. "No, unless the subject is formulated that way. Instead of calling it the Study of English Literature the course name should then be changed to something like Study of English Literature – Second Language or Non-native speakers".
- b. "No. I think that it does so to the extent that is appropriate for an English Literature module
- c. Not sure. This would depend on whether this means a lowering of the standard. If there is to be a lower standard, then 'no' those should not be taken into account".
- d. "No. A standardised scale means a qualification without compromise/perceived to be on par with competing universities".

Unsure –

"Unsure - learning should as far as possible be equitable and fair".

8. Question asked

If you answered Number 7 (Number 6 for tutors) in the affirmative, do you think that the scale adequately reflects the distance learning context?

Results

The majority disagreed

Tutors: 5 (71.4% of those who answered the previous question in the affirmative)

Markers": 5 (100% of those who answered the previous question in the affirmative)

Representative comments**Yes**

- a. "Where it says meaning is slightly obscured or not obscured. This is to say the standard of language is different but meaning is not obscured".

No

- a. "The current scale is designed in such a way to assess a homogenised student [body] and does not take into account specificities. As suggested above, a scale of 25 language, 25 structure, and 50 analysis could amend the disparity".
- b. "In my opinion, distance learning is less of the issue. What is more prominent is the context of situation and culture derived from the particular destination that distance refers to. This is covered by cultural and linguistic diversity as outlined above".
- c. "Could give additional focus on rewarding retention of skill taught on the course rather than skills which may have been imported by more privileged schooling".
- d. "Broaden level 3, use language the students can understand, give more detailed analysis on assignment 1 feedback".

9. Question asked

If you could make one change to the current scale, what would that change be?

Representative comments

- a. "I would conflate vocabulary and language use and weigh 25 for language, 25 for structure, and 50 for analysis".
- b. "I would definitely remove the word 'shaky' and replace it with a more professional/academic term. I would also suggest a more descriptive word for "mechanics" – Grammar? Form? Structure?"
- c. "Add another descriptor for referencing and style".
- d. "Probably the weighting between content and organisation versus language".
- e. "Added level to allow a broader range of mark allocation".
- f. "The naming of the levels. Levels 3 and 4 seem named in a way to not cause offence, yet 'acceptable' (3) and 'poor' (4) are clearer and of an appropriate register".
- g. "The lack of details in the description of mark allocations".
- h. "Changing of the 50/50 weighting of marks to 33/33/33 based on content, language and structure".
- i. "I think it could be useful for an explanation to be included with some elements of the scale, for example, what constitutes as an 'idiom issue' perhaps, in order to create clarity and consistency among those who are assessing students".
- j. "I would remove vocabulary range; this doesn't affect the quality of the argument or the quality of the writing, unless it somehow makes meaning ambiguous".
- k. "Broaden level 3 into 2 levels".
- l. "Rather a landscape table that can be read from left to right, for example or have both grids on the same page (one below the other)".

10. Question asked

What feature in particular would you like the revised scale to retain?

Representative comments

- a. "The notes on content are especially useful".
- b. "A focus on both content and language".
- c. "Its layout".
- d. "Simplicity and easy comparability between the different levels in the criteria".

- e. "A strong focus on those at risk of failing the module".
- f. "The fact that the scale contains divisions such as 'vocabulary/language use/mechanics' is a helpful feature, as well as the scale descriptions which indicate whether a student might be considered 'at risk' – this is a helpful feature to both students and facilitators in my opinion."
- g. "The broad ranges, allowing for the marker to use their discretion".
- h. "Level 1 and 2 are fine".
- i. "I like that content and organisation are not graded separately, as it provides some leeway for the marker's discretion".
- j. "The balance between how the content is conveyed and just interpreting the topic".
- k. "Most of it is very useful and the explanations are pertinent".

Other comments

- a. "I think that overly detailed or rigid rating scales are problematic. Simple scales (or rubrics) rely on the expertise of a marker/moderator, and will ensure there are efficient marking processes in place and fair marking in practice".
- b. "The rating scale is quite good until it gets to level 3 then misinterpretation can easily set in. The percentage of 54% to 32% is too big of a gap. It goes from fair to shaky but the criteria are all negative, how can it be interpreted as fair?"

6.5 OBSERVATIONS ON QUESTIONNAIRE FEEDBACK

The following observations summarise the opinions of markers and tutors as shown in the responses to the questionnaires.

6.5.1 Number of levels

As can be seen, the majority of Unisa markers and tutors commented on the need for more levels, in particular at Level 3, with its wide range (32% - 54%) which might cause difficulties in distinguishing between a pass and a failure mark. The majority of markers and tutors agreed that Level 3 should be divided into 2 levels to indicate the difference between "pass" (50% and above) and "fail" (below 50%). This was supported by the feedback from markers employed to mark the selected scripts for the current project. This issue was considered to be a priority, as the current, wide range could have potential consequences for the student because it offers no guidelines as to the cut-off point between "pass" and "fail". This large category is also not conducive to formative assessment as the difference between 54% and 32% it is too large, and students could become confused as to the reason for a script being classified as "pass" or "fail" when they are relegated to the same level.

There were also suggestions that the top end of the scale (Level 1) should be divided into 2 levels to encourage students to strive higher than 75% and prevent the possibly subconscious impression on the part of markers as well as students that “75% is the new 100%”. Although very few students achieve this high mark, the researcher concurred that adding a category at the top of the scale would “raise the bar”, especially in the case of students emerging from an educational system where 30% is considered to be a “pass” in some circumstances. It can be argued that these low expectations encourage a concomitant under-achievement on the part of many students.

6.5.2 Weighting of marks between content and language use

There were differing opinions about the 50/50 weighting of the marks between organisation/content and language use with those markers in favour of a change in weighting suggesting a lower weighting for language, given the specific target group and context. A different scale, which combines the two outcomes, could also be considered in order to rectify this matter.

Various weighting options were suggested. Of these, the consensus seemed to be that content should carry a greater weighting than language. However, those in favour of the *status quo* pointed out that a university course should not be too lenient with regard to “very faulty, non-academic language”. On the other hand, some responses indicated that some accommodation should be made for features of South African English varieties, as can be seen in the following remark:

- “What I think is useful to take into account is the student’s ability to analyse a text based on how they are instructed to do so. As in any other module, there is content for the student to learn. In this module, we teach them how to analyse in a particular way and therefore, ideally, they should be able to do so despite diverse backgrounds”.

These comments are also relevant to the issue of multiculturalism, and multilingualism.6.5.5

6.5.3 The extent to which markers adhered to the marking scale

Markers were guided in differing degrees by the rating scale, but few adhered to it rigidly at all times. Most markers used the scale as a guideline and the degree to which they adhered to it depended on circumstances such as time constraints (i.e. a tight marking schedule) or the characteristics of individual scripts that did not strictly reflect the norm (in which case the marker would use his/her own discretion, but would bear the scale in mind as a rough guideline). This evidence indicates the need for a scale that combines accuracy with ‘user-friendliness’. Thus, a fine line would have to be negotiated between validity (which is of paramount importance) and practicability.

6.5.4 The extent to which the scale should take the distance learning context into account

In this context, the importance of formative assessment, expressed in terms of understandable feedback, was stressed. This could mitigate the lack of face-to-face contact between marker and student. Most participants felt that more detail and/or clearer descriptions in the marking grid could assist students who rely on this feedback in the absence of face-to-face interaction. Furthermore, more levels might help students in obtaining better understanding of the features of their assignments that need improvement. If students are familiarised with the scale and can use it to understand the shortcomings (and positive features) of their assignments, the lack of face-to-face contact could be ameliorated. This could be achieved by including the grid in the marking feedback, for example, and exposing students to the scale in various media and forums, not only in the study material. It is also important to ensure that the terms used on the scale are understandable. The study material could be used for this purpose, and use can be made of social media and student platforms such as the e-tutor programme. For instance, a live-streaming session could be dedicated to explaining the marking grid to the students. These suggestions are explored further in Chapters 7 and 8.

6.5.5 The extent to which the scale should take the multicultural and multilinguistic target market into account

Although most agreed with the statement, some concerns were raised that the standards might be lowered. On the other hand, there was evidence of tolerance and

accommodation of varieties of South African English (SAE). Another issue to be considered is how factors such as unequal education, poor schooling and socio-political redress in the South African environment can be balanced with international exigencies in order to ensure that students can compete in a wider context. There was thus a problem with the implementation of this principle. This important question was discussed at other stages of the current research, particularly in connection with the design and construction of the alternative scales (Chapter 7). An issue that arises is the difference between descriptive and prescriptive views on grammar, and to what extent a prescriptive or descriptive approach is applicable in the South African context (Görlach, 1998; Silva 1997; Schneider 2011). This question should be considered against the background of the challenges of the socio-political and socio-economic factors relating to previously unequal schooling and also to the continuing poor education in poor socio-economic areas. The viability of addressing these issues through the design and use of a rating scale was questioned, and, if indeed possible, a further question was how it should be achieved. It was noted that the major criterion for language usage in the existing scale was intelligibility (i.e. to what extent meaning was obscured), and, although this could be interpreted differently by markers, it was believed that there was sufficient consensus to retain this as a yardstick. In addition, there would appear to be a growing acceptance of features of SAE, as evinced in the markers' comments (Section 3.5.2) and an awareness of the multicultural and multilingual complexity of the ODL target market. This apparently growing acceptance is supported by the following remarks by Unisa markers:

- “In my opinion, distance learning is less of the issue. What is more prominent is the context of situation and culture derived from the particular destination that distance refers to. This is covered by cultural and linguistic diversity”.
- “[The scale] [c]ould give additional focus on rewarding retention of skills taught on the course”.

Ideally, this debate should be initiated at the level of syllabus design and academic development, and then filtered down to be reflected in the assessment scale. The challenge to the scale developer is to design a scale that is accurate and balanced (i.e. measures what it is supposed to measure), as well as being user-friendly. Furthermore, designing the scale as a “quick fix” should be avoided while, at the same time

facilitating a fast and accurate turn-around time to meet the demands of the complex ODL administrative and academic environment. In addition, it should provide the students with enough formative feedback to enable them to improve their marks in future assignments and, ultimately, in the final examination (summative assessment). This was indeed a daunting task, but it was considered to be worth the attempt.

6.5.6 The perceived subjectivity of the scale

Although participants acknowledged that some features were open to subjective interpretation, they felt that this was inevitable when assessing an English assignment and could be remedied by training and consultation. A further factor was the large range represented by the levels, particularly Levels 1, 2 and 3. This could lead to subjective marking as the boundaries were unclear.

6.6 CONCLUSION

In this chapter, the results from the quantitative and qualitative methods employed in this research to validate and evaluate the existing rating scale have been presented in order to provide answers to the following research questions. The findings have been summarised briefly after each question.

1. What do the results of the empirical research process reveal about the validity of the existing scale?

- **Findings**

The scale was tested statistically and demonstrated good to satisfactory results, particularly as regards scoring validity. There was only one case of serious inter-rater discrepancy, but this was not sufficient to invalidate the scale. A subsequent calculation, excluding the marker concerned, showed an acceptable reliability rate (0.79).

Qualitatively, the scale demonstrated some weaknesses as regards construct validity, particularly in respect of Level 3, where the range was too wide and

could have a negative effect on marking accuracy. This, in turn, would also affect consequential validity, as an inaccurate failure mark could result in the student's failing to complete the module, and thus delay or even prevent the obtaining of a qualification, resulting in negative economic (and possibly psychological) consequences. Lack of directives regarding plagiarism also affected the validity of the existing scale negatively, especially since there is extensive evidence of plagiarism in the ODL context.

2. What are the observations of the tutors and markers who use the existing scale to assess examinations and assignments for this module?

- **Findings**

Concerns included the number of levels and, in particular, the range of Level 3. This was considered to foster subjectivity and therefore prevent accurate marking. Other issues raised were the weighting of marks for content and language use, and the potential subjectivity of certain guidelines (e.g. "shaky"). The importance of the scale in formative assessment was emphasised. This is a main issue, also discussed by markers of the pilot and main studies of the current research. Furthermore, plagiarism is a problem which is exacerbated in the ODL context, and has a negative effect on validity. The current marking grid makes no allowance for penalties with regard to plagiarism, and this was an area of concern to markers. It should be noted that plagiarism affects validity as the lack of a penalty results in an inaccurate or "false" mark.

3. What effect, if any, does the distance learning, multilingual and multicultural context have on the perceived and actual validity of the scale?

- **Findings**

There is minimal face-to-face interaction between markers and students. Thus, the scale should be clear and accessible to students as this is the main instrument of formative assessment. This should be borne in mind when designing a new scale. A close link was found between the challenges of ODL and those presented by the diverse multicultural and multilingual population group of this

research. It could be argued that the ODL context exacerbates these problems. While the scale might be measurably valid, it should be seen to be convincingly so by the target group and, by implication, should lend itself to accessible formative assessment. Markers should also be able to award marks that reflect the underlying abilities fairly. Factors affecting validity negatively were the unacceptably wide range of some levels (particularly Level 3) and the lack of penalty for plagiarism. Terms like “shaky” were perceived to be unclear and subjective.

The data generated by the quantitative processes and the information extracted from the qualitative research formed the basis of the next phase of the project, namely the design and testing of a new scale. This is discussed in the following chapter.

This chapter provided the bulk of the empirical research pertaining to the current scale, and progressed from an account of the pilot project to a detailed description of the quantitative process employed in the main study. This was followed by the description and results of the qualitative aspects of the research. The chapter concluded with an overview of the findings, presented in the form of answers to three of the questions posed by the study.

CHAPTER 7: DEVELOPMENT OF THE NEW SCALE

7.1 INTRODUCTION

This chapter contains a description of the process followed in this study to develop new scales, based on the quantitative and qualitative findings discussed in Chapter 6. The panel discussions are summarised and the stages of developing a new scale (namely, the design stage, the construction stage and the trial stage) have been discussed in detail. Evidence includes quantitative elements, in the form of statistical analysis, as well as the qualitative features extracted from the comments of markers employed at various stages of the process. Finally, the new scale has been presented and reasons given for this choice. The chapter concludes with a summary of the process followed and the results of this process.

7.2 Design and construction of the proposed new scales

During the process of revising and refining the assessment scales, use was made of the information gathered from the analysis of the data relating to the scale that was currently in use. This process was initiated by a panel of five experts, comprising experienced educators and examiners of English at Unisa and other institutions, as described in Section 5.8.1.3. Communication was conducted by means of face-to-face interaction as well as by means of electronic media such as email and Skype.

McNamara (1996) and Taylor (2002) agree that developing or reviewing an assessment scale includes three stages, namely: the design stage, the construction stage and a trial stage. This process was followed in the current research. Initial designs were interrogated during the design and construction phases, and were revised as a result of the trial stage, leading to constant revisions and interaction between the three phases. Although the process appears to be linear, it was, in effect, a recursive model.

Prior to the first workshop, described in Section 6.3.4, panel members were briefed about the aims of the project and provided with a copy of the research proposal for this

study as well as comments on the existing scale provided by the markers. This preparation assisted the panel in discussing the efficacy of the existing rating scale, and the design of alternative scales.

During this workshop, the results of the quantitative and qualitative data pertaining to the existing scale were discussed in order to determine the type of scale that would be most suitable in assessing the construct in the given context. The two-fold construct for the module was analysed in order to provide directives on the rating scale that reflected these constructs.

As described in Section 5.6.2 and Appendix C, the outcomes for the module were as follows:

Students credited with this module will be able to apply appropriate reading strategies to a wide variety of literary and non-literary texts in English. They will also be able to demonstrate basic skills of writing academic English.

The second specific outcome (“Demonstrate basic awareness of the creative choices made by writers of literary texts in English”), leads to the assessment criteria as follows:

- The dimensions of artistry and contrivance in the composition of literary texts in English are explored and explained through acceptable academic discourse.
- Accepted conventions of academic discourse are applied.

It was argued (Section 2.6) that wording such as “the dimensions of artistry and contrivance” and “acceptable academic discourse” is vague and would need to be clarified to prevent misinterpretation and subjectivity, and to facilitate the alignment of the assessment scale to the criteria. The outcome itself should be revised in order to reflect the current student demographic, in contrast to the past predominantly L1 student population. The dual purpose of assessment, namely its formative and summative functions, is nevertheless clearly stated in the outcomes, and this underlines the importance of valid scoring criteria that can be used for both of these functions.

In summary, the two areas covered by these outcomes are the ability to:

- read a range of texts with insight and discernment, as well as to be able to identify and discuss stylistic and technical features of the text; and
- write about these texts using basic academic discourse (the examination marker’s guidelines simplify this further by describing the discourse as “correct, standard English”).

It was decided to base the definition of the construct on these terms for the sake of the current research although, in practice, more clarity should be desired in terms of the definitions given in the outcomes. An additional problem was the degree to which the construct was reflected in the current scale, especially given the generic nature of the scale. For example, the Marking Grid of ENG1501 is identical to that of *English for Academic Purposes* (ENN1013F), although the NQF outcomes are different for ENG1501.

Furthermore, as can be seen in the extract from the existing marking grid, the criteria extrapolated from the NQF outcomes are mentioned parenthetically in Level 1, but not reiterated or specified at the other levels. This could lead to cross-referencing by markers, although it was conceded that the summaries provided at these levels did give some guidance. The panel also questioned the alignment of the criteria to the stated outcome (i.e. “[t]he dimensions of artistry and contrivance in the composition of literary texts in English are explored and explained through acceptable academic discourse”). The difficulty of alignment was understandable, given the vague wording of the outcome but, despite this concession, the panel felt that reference to the NQF outcomes was not sufficiently emphasised in the existing scale and gave the impression to panel members that the description had been added (or “tacked on”, as one member put it), and, in fact, it was discovered that the reference to the NQF outcomes was itself generic, as demonstrated by Table 7.1. In short, the grid and assessment criteria have not been adequately revised to reflect the changes that have taken place in the student body.

Table 7.1: Extracts from existing marking grids for ENG 1502 and ENN 1013F

Score	Level	Criteria
25-19 (100%-76%)	1 Excellent to very good	<p>ENG1501 Content: focused on assigned topic, thoroughly developed, clearly demonstrating the skills required by the NQF criteria (e.g. familiarity with - recognising and recalling - the subject matter; understanding it; application of this information; analysis, for instance of relationships; evaluation, for example critiquing different approaches) Organisation: generating a piece of writing (such as an essay) with ideas clearly stated, succinct, well-organised, logically sequenced, cohesive, well-supported</p> <p>ENN1013F Content: focused on assigned topic, thoroughly developed, clearly demonstrating the skills required by the NQF criteria (e.g. familiarity with - recognising and recalling - the subject matter; understanding it; application of this information; analysis, for instance of relationships; evaluation, for example critiquing different approaches) Organisation: generating a piece of writing (such as an essay) with ideas clearly stated, succinct, well-organised, logically sequenced, cohesive, well-supported</p>
18-14 (75%-56%)	2 Good to average	<p>ENG1501 Content: fairly sound demonstration of skills, mostly relevant to topic, lacks detail Organisation: loosely organised, logical but incomplete sequencing and signposting</p> <p>ENN1013F Content: fairly sound demonstration of skills, mostly relevant to topic, lacks detail Organisation: loosely organised, logical but incomplete sequencing and signposting</p>

Source: (Tutorial Letter 101/1, 2013: 30)

The generic approach demonstrated in Table 7.1 was not considered to be satisfactory, especially in the more specialised context of literary studies which require specific skills such as interpretation, substantiation with reference to the text, and the understanding and appreciation of imagery. This literary content is not relevant to the aims of ENN 1013F, which was designed to teach writing skills and, therefore, was not literature-based. Thus, the construct validity of the existing scale for ENG1501 was once again interrogated.

Furthermore, although the relationship between the two main outcomes, which the existing scale reflects as content/organisation and language use, is not clearly stated, the panel argued that, when assessing literary assignments, knowledge of, and insight into, the literary text should take precedence over language and style. This would obviate the possibility that a well-written script that demonstrates minimal knowledge of the literary text could be awarded a pass mark. This potential difficulty had been pointed out by various markers (Sections 6.3.3, 6.4.4 and 6.4.5), typified by the following remark made by a marker:

For me the most important point in the rubric should be ‘Does the candidate address the question?’ Despite grammar errors, the question is whether the candidate can construct and develop an argument and get it across.

In the light of these observations, it was agreed that a solution should be found for this eventuality. It was felt that priority should be given to finding a satisfactory weighting between content and language. This could possibly take the form of a combined or unified mark, such as that shown in a two-dimensional scale.

The panel was also made aware of the importance of employing the scale for formative assessment, particularly in the complex multicultural and multilingual environment of South African distance learning. This was in addition to the scale’s obvious importance in summative assessment. As McKenna (2007: 25) points out: “Rubrics function as feedback forms for learners by identifying areas of the assessment where the learner has not met a stated criterion”. McKenna (2007: 25) observes further that frequently feedback is “given on the rubric” and augmented by comments in the margin and at the end of the rubric. Unlike the participants in an “ideal situation” envisaged by McKenna (2007: 25), students currently registered for ENG1501 do not have the opportunity to resubmit their assignments using the feedback on the rubric as a “developmental tool” (Spencer 2009: 103), but it is relevant to the present study that lecturers in McKenna’s research found that “the reworking of the assignment was greatly improved when problem areas were pointed out through a rubric” (McKenna 2007: 25). In the case of the target group, Spencer (2009: 103) points out that:

Sadly, with the... introduction of semester modules rather than year courses, there is no longer adequate turn-around time in this distance teaching context to require a resubmission assignment.

The introduction of a year-long course to replace the semester module ENG1501 in 2020 should help to mitigate the problem of turn-around time, although deadlines will still be tight, as more assignments will be prescribed. The importance of the rubric for formative assessment remains, and in fact the envisaged longer course could encourage interaction with the assessment scale, especially if a re-submission assignment is introduced, as was attempted by Spencer (2009), whose research was conducted in this teaching environment. A suggestion was made that the assessment scale could be included with the scripts (in the same way that the declaration of originality is submitted) and that markers could use this as a feedback tool, by underlining or ticking relevant features (those in need of attention and/or those demonstrating good work). The scale could then be returned to students along with the marked script and any other comments that the marker deems necessary to make (McKenna 2007: 25). This would provide an effective and relatively easy way of commenting on the student's work, given the volume of scripts that each marker is expected to mark, and the timeframe.

The panel agreed with the views on empowering students in an ODL context, and that using the assessment scale as a training tool would enable the students to "internalise standards and independently judge their work against the listed criteria" (Spencer 2009: 105). Thus, as Spencer (2009: 105) points out, guided self-monitoring benefits the marker, by facilitating marking, as well as the student, who will gain by being encouraged to engage with the rating scale in order to improve their writing style and/or content knowledge and interpretation. Spencer (2009: 105) observes that this practice could encourage progress "towards personalised, individualised learning", in the challenging ODL environment in which influxes of thousands of scripts, identified only by the student's name, number and contact details, are processed.

In connection with the value of the current scale in the area of formative assessment, the panel considered comments made by tutors and markers. An example of these comments was:

While I see the sense in using [the] scale as a marker... I don't think it provides enough information for the student. There is a big difference in marks between 56% and 75%, for example, and the marking grid does not explain sufficiently the difference between an Average assignment and a Good assignment. I think more detail could be helpful here, so the student can have a clearer idea on how to elevate an Average mark to a Good mark, for example.

At this stage, the panel decided that the importance of formative feedback would be discussed in more detail during the scale design and development stages and included in the final recommendations. The panel also briefly discussed the inclusion of graphics in the rating scale and the use of technology (such as cell phones and social media). It was agreed that electronic media could provide a useful means of communication with ODL students, especially if the marking grid were made available on these media as well as in the tutorial letters. Members were unsure about the effectiveness of graphics in the rating scale, citing possible different cultural implications and miscommunication. This was discussed more fully at a later stage of the process (Section 7.8).

While discussing the current scale (Sections 6.3.4. and 6.4.5), the panel analysed selected examples of student writing in order to identify the salient features that distinguish performances at different levels of proficiency (Hattingh 2009: 172-173). These scripts had been chosen by the researcher to represent all four levels of the existing marking grid, and were sent to the panel members electronically prior to the workshop, although no marks were provided. At the meeting, marks for the scripts were suggested and reasons given for the marks allocated. The existing scale was then examined closely to determine whether it met the necessary criteria for content/organisation and language use (as indicated by the outcomes given in Appendix C and summarised in Section 7.2), and how the scale could be improved, revised or replaced. At the subsequent discussion to design alternatives to the existing scale, other scales were surveyed with a view to incorporating features that the panel considered to be relevant. These included: the multi-trait scale recommended by Hattingh (2009: 276), the grid used for the De Beers English Olympiad, the grid used by the English Department of the Nelson Mandela University, and other grids used by the Department of Education, incorporated in the discussion on formative assessment (7.8).

Having discussed the information gathered from the analysis of the data, the panel commenced the design and construction of alternative scales. This process was continued at two subsequent meetings and by means of electronic and telephonic interaction between panel members.

7.2.1 The design of the scale

Suggested changes were made according to the evidence produced by the empirical research on the existing scale. The comments and recommendations of the markers and tutors were taken into account and these formed the basis of the design stage. The changes recommended have been presented thematically below:

7.2.1.1 *Number of levels*

The panel was unanimous that Level 3 of the existing scale should be divided into two levels in order to distinguish between a “pass” and a “fail” mark. This was altered immediately as it was considered to be the most glaring shortcoming of the existing scale. Therefore, it was agreed initially that there should be five levels instead of four, although this decision was modified at a later stage, when the possibility of creating another category at the top end of the scale was examined. The discussion took into account the feedback from markers and tutors, encapsulated in the comment that “75% has become the new 100 %”, an observation corroborated by the panel members. Thus, after some debate, it was agreed to add an ‘Exceptional’ category as Level 1.

Initially, a suggestion was made to eliminate the second lowest category (“at risk” or “shaky”) in order to streamline the scale. However, ultimately, it was decided to retain this category because of its value in formative assessment. Students need to know the reasons for failure, and those in this borderline category could be guided by the stated criteria given in the scale. The scale was thus amended to include six provisional categories as shown in Table 7.2 below.

Table 7.2: Extract showing amended categories

Assessment ENG1501	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
Classification	Exceptional (High Distinction)	Excellent (Distinction)	Good to Above Average	Average to below Average	FAIL: Borderline	FAIL
Mark out of 25	25-22 or 100-85%	21-19 or 84- 75%	18-15 Or 74- 60%	14-12.5. or 59-50%	11.5--9 or 49-36%	8-0 or 35- 0%

7.2.1.2 Type of scale

The type of scale was also discussed with panel members, who examined three types of scale, namely:

- The Likert 7- point scale;
- A holistic scale similar to the existing scale, but with significant modifications; and
- A two-dimensional grid, similar to those employed by the English Department of Nelson Mandela University and the De Beers English Olympiad.

These scales were then discussed individually as follows:

(a) The Likert Scale

The Likert Scale favoured by Hattingh (2009: 190) was examined and discussed at length. In a typical Likert Scale such as that proposed by Hattingh (2009:190), and shown in Table 7.3, the categories (micro-categories as well as macro-categories) are based on the features identified by the construct and/or by examples of student writing. Markers assign marks to individual features before calculating the final mark (i.e. the total of the marks allocated to the individual features). The scale makes use of bi-polar descriptors, which describe only the extreme ends of the scale for the individual feature. Hattingh (2009: 191) claims that these characteristics of the scale reduce the possibility of teachers assigning impression marks to an essay. The final draft of the scale is shown in Table 7.3 (Hattingh, 2009: 190)

Table 7.3: Final draft of proposed Likert Scale assessment grid

FINAL PROPOSED DRAFT RATING SCALE											
	Poor			Adequate			Very good				
	0	1	2	3	4	5	6	7			
A. CONTENT											
1. No insight into and understanding of topic.	0	1	2	3	4	5	6	7	Outstanding insight into and comprehensive understanding of topic.		
2. Hardly any originality and/or little interest/ mundane.	0	1	2	3	4	5	6	7	Highly original/ Fresh perspective/ original/ engaging creativity.		
3. Irrelevant and immature ideas	0	1	2	3	4	5	6	7	Mature and thought provoking ideas.		
4. Does not follow the conventions of essay type.	0	1	2	3	4	5	6	7	Ideally follows conventions of essay type.		
5. Incoherent flow of ideas.	0	1	2	3	4	5	6	7	Highly coherent flow of ideas.		
B. STRUCTURE AND STYLE											
6. No division into introduction, body, conclusion.	0	1	2	3	4	5	6	7	Effective division into introduction, body and conclusion.		
7. No paragraphing.	0	1	2	3	4	5	6	7	Effective paragraphing.		
C. GRAMMAR											
8. Incorrect syntax.	0	1	2	3	4	5	6	7	Correct syntax.		
9. Incorrect tense & concord.	0	1	2	3	4	5	6	7	Correct tense & concord.		
10. No variety in range of sentence types.	0	1	2	3	4	5	6	7	Wide variety in range of sentence type.		
11. Multiple errors in spelling & punctuation.	0	1	2	3	4	5	6	7	Error-free spelling & punctuation.		
D. VOCABULARY											
12. Limited range.	0	1	2	3	4	5	6	7	Extended range.		
13. Inappropriate style, diction & register.	0	1	2	3	4	5	6	7	Highly appropriate style, diction & register.		
14. Ineffective use of linking devices (words & phrases).	0	1	2	3	4	5	6	7	Sophisticated use of linking devices (words & phrases).		
E. LENGTH											
Deviation from requirement.	0	1	2						Adheres to requirement.		
TOTAL									100/2		

Source: Adapted from (Hattingh 2009: 190)

Hattingh (2009: 191) states that “the use of bi-polar descriptors... reduces ambiguity in descriptors and should improve rater consistency”. Hattingh also believes that the scale has the added advantage of providing diagnostic detail on learners’ abilities, thus fostering meaningful formative assessment.

However, the disadvantage of the scale is that it is time-consuming as markers have to add up a number of scores to obtain the final assessment. This would present a particular problem in the field of distance education, and specifically in the pressured marking environment that frequently prevails at Unisa, owing to the large number of students and the often extremely tight deadlines described in Section 6.3.2. Thus, while the panel acknowledged that practical convenience should not be prioritised at the expense of a comprehensive construct representation (Weir 2005: 49), the belief was

that the scale should also be designed to be as practicable as possible, especially in the context of Unisa.

Furthermore, while the panel members also acknowledged that the holistic nature of the existing scale could lead to subjectivity and possible misinterpretation by markers and students, they disagreed with the criticism of Hattingh (2009: 190) that:

...scales such as Jacobs *et al.* (1981)... may lead raters, consciously or not, to concentrate more on one micro-feature rather than consider all features related to a criterion equally. Raters may thus equate a criterion with one salient feature instead of a collection of related features, which is unfair towards test takers who may be more developed in another related feature.

This problem did not emerge in the statistical findings on inter-rater variance for the existing scale and was not cited as a major difficulty in the qualitative evidence, although the possibility of subjectivity was mentioned in certain of the questionnaire answers. The panel believed that thorough briefing and training of markers, combined with a well-organised system of moderation, could obviate the potential problem of subjectivity. A benchmarking exercise would clarify criteria and descriptors found in the rating scale. This is already standard practice for markers of ENG1501. The grid should also include an instruction to markers to avoid adopting a “checklist” approach to the marking process

In summary, while it was agreed that practicability (or ‘user-friendliness’) should not be a predominant factor (Weir 2005: 49; Hattingh 2009: 193), the combination of accuracy and ease of use should be considered as an ideal, particularly in the time-constrained distance learning environment. The Likert Scale proposed by Hattingh (2009: 109) was considered time-consuming and unwieldy (or, as one panelist put it, “too fiddly”), and there were doubts whether it would lend itself to the context. It was pointed out that a grid that is not user friendly would not be viable in the pressurised ODL teaching context in which a marker might be required to assess 200 scripts per week. Furthermore, in this pressurised environment, the Likert Scale could also be open to marker error if marks are not added accurately.

It was also pointed out that, ironically, a scale similar to that proposed by Hattingh (2009: 109) could be counter-productive because it might lead to concentration on individual features rather than consideration of how these interact to form an overall view of the student's competence. This was reflected in tutors' and markers' comments in the current study, such as: "It seems strange to give... separate marks for an essay instead of looking at how all the criteria work together". These remarks by stakeholders seem to favour an even more integrated approach than that of the existing scale and, thus, by implication, led to further reservations about the suitability of the more fragmented Likert Scale for the context of the present research.

It is noted that, while the participants in Hattingh's research generally favoured the Likert Scale, subjectivity was still considered an issue when scoring according to this scale (Hattingh 2009: 225). For instance, one marker commented that, in scripts lacking cohesion and coherence, "it is going to be difficult for one to allocate marks for the learners", and another asked, "How good is 'very good'?" and added that "The wording/descriptor of each level should perhaps change.... [There is] no such [thing] as poor – how poor is a thing?"

Although Hattingh (2009: 225) does address these issues, the fact remains that it is almost impossible to avoid subjectivity and it could be argued that more descriptors to describe each level and category would be more helpful to markers and students. The bi-polar Likert Scale is not conducive to this, as its inclusion could make the scale more unwieldy.

Furthermore, the difference between the ENG1501 target group and that of Hattingh (2009) should be taken into account. In the case of ENG1501, knowledge, interpretation and appreciation of a specific body of content (the prescribed literary texts) are required. By contrast, the construct upon which the research by Hattingh (2009) was based emphasised writing ability at a National Senior Certificate (English First Additional Language) level. The panel was of the opinion that a "checklist approach" should be avoided in the present context. In other words, markers should not mark the features in isolation (as a "checklist") but should see them rather in interaction with one another. Thus, a script that does not meet every criterion for a level will not necessarily be penalised by "demoting" it to the lower level. However, it was agreed that, in order to

promote validity when marking globally, criteria could be stated more clearly than those described in the existing scale. Furthermore, it was recommended that markers of ENG1501 be informed of this global approach by means of an instruction on the marking grid. This could be reinforced during the training sessions that take place at the beginning of each semester.

The following examination guideline sent to examination markers of ENG1501 reinforced these viewpoints in favour of a global approach to script marking:

As a simple guide, award high marks for answers that are well substantiated with evidence from the texts. Award a mark of 24 (out of 50) for borderline cases; and 21 or 22 for those you feel could be allowed a second chance in the supplementary exam. Definite fails should receive 19 and below.

The existing grid is used mainly as a guideline by well-trained and experienced markers and used only in cases where markers are uncertain about a script. However, the comment on which this finding was based was made in connection with summative assessment and that more detailed marking would be required to provide formative feedback.

The panel then considered other possible scale designs, namely, a revised version of the existing scale and a two-dimensional scale.

(b) Revised version of existing scale (Model 1)

A revised version of the existing scale, incorporating features from various scales, including the Likert Scale, was suggested. This led to the construction of an experimental model based on the layout of the existing scale, but landscaped to include both content/organisation and language use, one below the other (or side by side), on one page for ease of reference. Table 7.4 below shows the Content/organisation section of the grid.

Table 7.4: Extract from revised grid (Model 1)

Classification →	Exceptional (Distinction)	Excellent (Distinction)	Good to Above Average	Average	'Borderline' FAIL	Seriously at risk: Fail
Mark →	25-22	22-19	18-15	14-12.5	11.5-9	8-0
1. Content/organisation. Criteria a. Insight: To what extent does the answer show maturity, understanding and originality? b. Awareness of stylistic/ technical features: Are these accurately demonstrated? c. Substantiation: Is the answer supported by appropriate reference to the text? d. Relevance: To what extent has all relevant information been included? Has the question been fully answered? e. Coherence – Does the answer flow together? NB MARK GLOBALLY	a. Original, sensitive, mature interpretation and insight. b. exceptional, original and sensitive c. Unfailingly well-supported, shows depth and insight d. Extremely relevant, well chosen, valid ideas, all points fully covered e. Shows exceptional focus, cohesion, seamless organisation.	a. Thorough, incisive, original b. Excellent, good examples. c. Extremely well supported with apt examples. d. Extremely relevant, all points covered. e. Exceptionally well structured, focused, coherent.	a. Good grasp of issues, some originality b. Well-demonstrated appreciation. c. Generally well substantiated. d. Mostly relevant, most issues addressed, points of question covered e. Well organised and coherent.	a. Adequate, lacks depth, little originality b. Features occasionally discussed, usually correct c. Some substantiation, but mainly just thoughts on the question. d. Fairly relevant, point sometimes missed. e. Loosely organised but still coherent.	a. Insight inadequate, little understanding of issues b. Features seldom discussed, shows lack of knowledge c. Not enough substance or relevance, ideas insufficiently supported d. Many statements lack relevance e. Ideas confused or disconnected, not enough logical sequencing or development, little signposting	a. Serious errors of understanding, extremely little evidence of knowledge of text b. Features ignored. c. Unsubstantiated. d. Irrelevant 'misses the point'. e. Incoherent, disjointed. Plagiarised OR Not enough to evaluate.
	SUMMARY Exceptional insight and organisation.	SUMMARY: Comprehensive, original, apt and mature insight, excellent organisation.	SUMMARY: Sufficient understanding well organised, comprehensive, relevant	SUMMARY: Adequate understanding lacks originality and depth	SUMMARY Frequently irrelevant, poor organisation	SUMMARY: serious errors, irrelevant, confused. OR Plagiarised.
Sub-total 25 marks						

The grid shown in Table 7.4 was considered to be more user-friendly than the Likert Scale. In addition, the panel agreed that the revised directives were closer to those of the outcomes of the module.

(c) The two-dimensional scale (Model 2)

The panel then considered designing a two-dimensional, matrix-style, experimental scale adapted from those of the De Beers English Olympiad, Nelson Mandela

University (NMU) and National Department of Education (Senior Certificate). These grids were discussed at some length to ascertain which of their features could be adopted or adapted to meet the requirements of the target group and the construct as described in the outcomes. The grids are shown in the tables below.

The De Beers English Olympiad grid is used for summative assessment, although the grid is provided as a guideline in the study material and gives comprehensive directives. Although some schools arrange voluntary classes for learners enrolled for the examination, Olympiad candidates rely chiefly on printed study material for guidelines on the prescribed texts. This is similar to the situation of the ENG1501 target group.

Table 7.5: Marking Guide: De Beers English Olympiad						
S T Y L E A N D L A N G U A G E						
C O N T E N T		A Excellent Style lively and crisp/Clipped/Excellent diction/Fluent/Phrasing and analogy relevant/Grammar/spelling/ punctuation/ paragraphing almost flawless/Use of quotation stylish and useful.	B Above average Diction appropriate and sometimes skilful. Academic. Very good grammar, spelling, paragraphing and punctuation. Knows how to introduce quotation stylishly.	C Average Good, useful diction. A fairly academic style, untainted by colloquialism and slang. Occasional succinct imagery. Spelling and punctuation errors do not detract. Paragraphing correct/Useful quotations	D Poor Diction sometimes inappropriate. Might lean towards colloquial usage Spelling and punctuation Paragraphing might be inconsistent/ Evidence of quotation but thin on frequency and depth/Language a detractor/Pedestrian	E Non-idiomatic Bad. Inconsistent spelling that detracts from meaning. Non-sentences/ Very difficult to follow/ Beginning to irritate/ Non-paragraphs or no paragraphs/Language a barrier
	1. Distinguished Candidate has answered very well. Subtle references – intelligent insight. Original, fresh. Literary skill distinguished. Coherent. Logical. Relevant. Substantiation: Intelligent. .	Silver + All criteria apply 45 – 50	Silver + Most Criteria apply 40 – 44	Silver Some criteria apply 35 – 39	Bronze Relevant/coherent/substantiated but pedestrian and/or beginning to lose relevance/coherence/substantiation 33-34	N/A
	2. Skilful Candidate has answered question well. Relevant referencing. Really good potential but missing real insight. Very competent/coherent/ Workmanlike/ Argument sound and delivered with confidence.	Silver+ All criteria apply 40 – 44	Silver Most criteria apply 36 – 39	Silver Some criteria apply 35	Bronze Language a detractor 32-33	N/A
	3. A Competent Answer Candidate has answered question fairly well. Some referencing. Shows potential, but missing insight. Competent/coherent/ Workmanlike/ Argument fairly sound.	Silver Most criteria apply 36-39	Silver Most criteria apply 35	Bronze Some criteria apply 33-34	Bronze Some criteria apply 30-32	
	4. An Acceptable Answer Some knowledge of the text and genre. Answer shows relevance. Ordinary/ predictable, boring/. could lack coherence	Bronze Has textual, character or thematic references and knows how to quote. Shows flashes of insight 33– 34	Bronze Has textual, character references and has an idea of theme. Familiar with text/most criteria apply 31 – 32	Bronze Some textual and character refs/ Shows grasp of a central theme. Some Gaps 29/30	Merit Acceptable content marred by language. 25-28	Participation Language is a barrier to meaning/Evidence of opinion but tainted by cliché/shallow 21 - 24
	5. Level of Competence partially acceptable according to: Relevance of substantiation and quotation, Focus on question, coherence Some understanding of text and genre	Merit Shows an appreciation of theme/is able to quote correctly/ Criteria appreciably applied 28 – 29	Merit Shows appreciation of the text/Most of criteria observed 26 – 27	Merit Shows an understanding of needs of question/some of criteria observed 25	Participation Has an opinion and some idea of the needs of the question/some criteria observed/marred by language errors/might be cliché/unoriginal 21 – 24	Participation Has an opinion on the topic/ question/Serious language flaws a barrier to meaning 20
	6. UNCLASSIFIED: Only one section / a little of each answered. Written far too little. Unable to classify					UNCLASSIFIED: Award a mark – less than 20 , depending on how much was written

Although the target group of the English Olympiad comprises High School learners, the relative lack of face-to-face interaction is similar to the ODL context. In the English Olympiad context, there is no explicit examination feedback, other than a final mark and ranking (indicated by Gold, Silver+, Silver, Bronze, Merit, Participation etc.). The grid is used to evaluate an essay-type question, and has been used with only slight alterations for several years, although there is a different examiner and examination topic each year. The candidates are identified by examination numbers only.

The panel was of the opinion that the De Beers Olympiad grid was comprehensive and that the two-dimensional format integrated the components of language use and content. Thus, the two-fold aims of the examination, namely the interpretation of the prescribed texts and the ability to express this fluently and intelligibly were addressed successfully. However, a suggestion was made that the grid should be “switched around” and that “Content” should replace “Style and language” on the horizontal plane of the grid. This would give it prominence, since the most important criterion should be content knowledge and insight.

Having examined the English Olympiad grid, the panel then turned its attention to another two-dimensional grid, this time in the domain of tertiary studies. This was the grid used to assess literature assignments and examinations at the Nelson Mandela University (NMU). The grid is shown in Table 7.6 below.

Table 7.6: Marking Guide: Department of English, Nelson Mandela University

AXES OF MARKING GUIDE								
EXPRESSION	CONTENT							
	1 Exceptionally insightful, logical and well- substantiated	2 Very clear understanding of main issues/ exceptional comprehensive- ness	3 Clarity of thought and flashes of insight/ well ordered and comprehen- sive grasp of main issues	4 Painstaki ng and thorough/ good grasp of main issues	5 Just covers the ground/ no evidence/ too general	6 Second- hand or trite and thin/ patchy relevance	7 Poor, confused/ irrelevant	8 Very poor/ right off the point
A Lucid and polished	85+	80+	75+					
B Clear and proficient	80+	78+	73+	70				
C Fluent and careful	78	75	72	65	58			
D Competent and correct	75	70	65	60	55	52	45	40
E Basically correct but some errors	70	68	62	58	52	50	42	35
F Flat, venial faults in grammar, clumsy expression		65	58	55	50	45	40	30
G Faulty grammar or idiom		58	55	50	45	40	35	25
H Very faulty grammar or idiom			45	40	37	35	30	10-

Panel members were impressed by the conciseness of the NMU grid and particularly the descriptions of the categories. These avoid standard descriptors such as “good” and ‘poor’, and most possibly promote quick, easy and accurate marking. However, the grid might not be suitable for formative assessment in a distance learning context where more commentary is necessary to guide the student, given the lack of face-to-face contact with the lecturers.

The panel agreed to develop a two-dimensional grid based on these examples, (with input from members of the examining body of the De Beers English Olympiad). It was decided to bear the directives of the English Olympiad grid in mind and to also consider the category descriptors of NMU.

Two panel members initially expressed preference for a scale that was similar in layout to the existing one. The reasons for this opinion were:

- As noted in Sections 6.3.3, 6.4.4 and 6.4.5, three of the markers of ENG1501 liked the existing scale and found it easy to use (e.g. “I like the layout”).
- The criticism of two-dimensional grids is expressed as follows by a participant in the research study of Hattingh (2009: 207):

The... scale requires raters to rate in two directions (vertically and horizontally). It is sometimes difficult to reach a crossing point that represents a fair mark on the current scale.

However, three panel members who had used both the NMU and De Beers Olympiad two-dimensional grids (as well as Departmental grids, praised them for their integrated approach, and claimed that, in their experience, the problem of rating in two directions was quickly overcome once the marker became familiar with the process. These panel members found these grids to be “user friendly” because they felt that they led to quick and accurate assessment. It was also noted that many tutors and markers of ENG1501 had not been exposed to, or questioned about, a possible two-dimensional scale, and that many comments had indicated a preference for integrated marking. Subsequently, those Unisa markers who were consulted (n = 4) expressed an interest in the two-dimensional scale and felt that it could be considered seriously as a viable alternative to the present scale. Furthermore, panelists and markers agreed that the grid would be conducive to formative assessment as it would provide the student with one integrated mark, as well as criteria for content/organisation and language use.

The panel thus decided to construct two rating scales as possible replacements for the existing scale. These would be a revised version of the existing scale (Model 1) and a two-dimensional grid (Model 2).

It was confirmed that both scales would be tested qualitatively and quantitatively, following a similar process to that followed in testing the existing scale (Chapter 6). A description of the construction and testing of the scales follows. Since a recursive model was adopted, the testing of each scale is described after the initial construction phases of the scale in question.

7.3 CONSTRUCTION OF THE REVISED VERSION OF THE EXISTING SCALE (MODEL 1)

The focus of the discussion on the revised scale was on the number of levels. The construction process then eliminated terms like “shaky” that had been criticised by markers as being too vague and subjective. Furthermore, a summary was provided at the bottom of each column in order to facilitate marking, while more detail was given in the other segments of the scale. The marker could refer to this detail if more clarity was needed in the case of an individual script and, importantly, could be used by the student for formative purposes.

In the first version of the “revised” scale, the range represented by each criterion is indicated on the left-hand side of the scale. For ease of reference and in order to clarify the descriptors at each level, the designers attempted to align the criteria with the descriptors at each level as shown in Table 7.7 below.

Table 7.7: Revised scale: Model 1 (Draft 1) Total 50 marks

Classification	Excellent (Distinction)	Good to Above Average	Average	Borderline FAIL	Fail
Mark	25-19	18-15	14-12.	11.9	8-0
1. Content/organisation. Criteria a. Insight: To what extent does the answer show maturity, understanding and originality? b. Awareness of stylistic/technical features: Are these accurately demonstrated? c. Substantiation: Is the answer supported by appropriate reference to the text? d. Relevance: To what extent has all relevant information been included? Has the question been fully answered? e. Coherence – Does the answer flow together? NB MARK GLOBALLY	a. Thorough, incisive, original b. Excellent, good examples. c. Extremely well supported with apt examples. d. Extremely relevant, all points covered. e. Exceptionally well structured, focused, coherent. SUMMARY: Mature, original, comprehensive. Shows 'sparkle'.	a. Good grasp of issues, some originality. b. Well-demonstrated appreciation. c. Well substantiated. d. Mostly relevant, most issues addressed, points of question covered. e. Well organised. SUMMARY: Sufficient understanding, well organised, comprehensive and relevant	a. Adequate, lacks depth, little originality b. Features occasionally discussed, usually correct c. Some substantiation, but mainly just thoughts on the question. d. Fairly relevant, point sometimes missed. e. Loosely organised but still coherent. SUMMARY: Adequate understanding, lacks originality and depth, misses some important points.	a. Insight Inadequate, little understanding of issues b. Features seldom discussed, shows lack of knowledge c. Not enough substance or relevance, insufficient support for ideas d. Many irrelevant statements e. Ideas confused or disconnected, not enough logical sequencing or development, little signposting SUMMARY: Frequently misses the point, disconnected, largely irrelevant	a. Serious errors. b. Features ignored. c. Unsubstantiated. d. Irrelevant 'misses the point'. e. Incoherent., disjointed. f. Plagiarised OR Not enough to evaluate. SUMMARY: serious errors, irrelevant, confused. OR Plagiarised.
Sub-total 25 marks					
Criteria 2. Language and style a. Vocabulary- is there a sufficient range of vocabulary and effective word choice? b. Register and tone- appropriate for academic writing or too colloquial or pretentious? c. Language errors – are errors negligible or are they frequent, <u>impeding meaning</u> ? NB MARK GLOBALLY	a. Sophisticated range, excellent word choice, b. Very appropriate. Clear, formal but not pretentious or verbose. c. Negligible language errors: SUMMARY: Clear, fluent, articulate	a. Good word choice and range. b. Generally appropriate. c. Some language errors, <u>but meaning not impeded.</u> SUMMARY: Clear despite errors	a. Small but adequate range, some errors of word choice. b. Occasional lapses (e.g. too colloquial or verbose) c. Frequent language errors, but <u>meaning seldom impeded.</u> SUMMARY: Adequate but pedestrian (dull)	a. Small range, frequent issues of word/idiom, choice, usage b. Inappropriate register. c. Frequent and serious language issues, <u>meaning confused or obscured</u> SUMMARY: Difficult to understand due to serious and distracting errors.	a – c Serious and distracting errors, barely intelligible. OR d. Not enough to evaluate. OR e. Plagiarised SUMMARY: Barely intelligible
Sub-total 25 marks					

7.4 TESTING OF THE REVISED SCALE (MODEL 1)

The planning and design process was followed by a trial of the revised scale during which the assignments were scored by five members of the previous group of markers in order to ensure reliability, and to evaluate the strengths and weaknesses of the scale. Five of the original group of markers were unavailable, but it was considered that the smaller group would be sufficiently representative. The revised scale was tested by means of following the same procedure as before:

- marking by panel members and markers of the sample of scripts (n = 60);
- statistical calibration of the results;
- revising the scale;
- refining the scale;
- repeating the process if necessary.

7.4.1 Marking of scripts

Five markers were employed for this process, and Scripts 1-60 were marked according to the new scale. As previously, all markers had post-graduate qualifications, a minimum of 10 years' experience in teaching in an L1 and L2 environment to English, Afrikaans, Xhosa, Sotho and Zulu-speaking students. Two markers were L2 speakers of English, and one spoke Xhosa as L2. The markers were briefed as previously (Section 6.3.4).

The results of this marking exercise were quantitatively and qualitatively analysed as in the testing of the existing marking grid in order to modify and refine the evaluation approach adopted for this research. Furthermore, markers were asked to submit comments that were discussed at each stage of the process. This provided rich information to reinforce the quantitative aspects of the data.

7.4.2 Quantitative analysis

The following quantitative results arose from the analysis. Figure 19 shows the results of measuring the variance of scores given by 5 markers, marking 60 scripts.

This map shows the markers clustered around the 0 logit mark, with Marker D awarding slightly higher scores than the rest, although still within an acceptable logit range. The “person” or script map clustered above the 0 logit mark, although there were 2 scripts situated below this mark, at -1 and -2 respectively. In general, however, the map suggests a higher ability level than that of other maps presented in this study. This higher level could be attributed to the relatively higher marks awarded to language use.

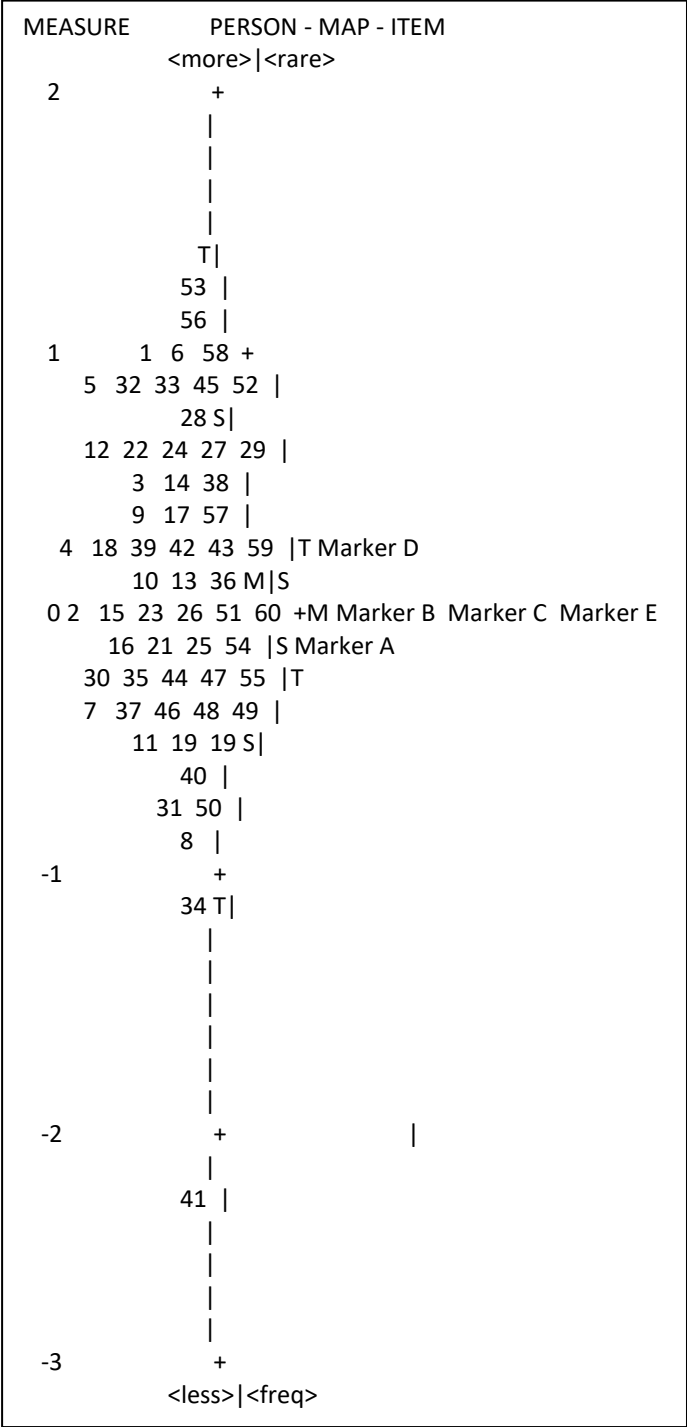


Figure 7.1: Model 1, 60 scripts, 5 markers

The scale was then tested for reliability as shown in Table 7.8.

Table 7.8: Reliability Model 1

Model	Script number	Raters	Script Reliability	Cronbach Alpha	Rater reliability
Model 1	Scripts 1 - 60	A-E	.97	.97	.95

The results indicated very good scoring validity (reliability), indicating high levels of inter-rater reliability and test “reproducibility”. The difficulty level and the range of marks also indicated that raters could distinguish the criteria given by the rating scale. This suggested a measure of construct validity, although there was concern that the relatively high marks awarded to language use had an effect on the construct validity of the final mark. The qualitative findings reflected in the comments and questionnaires were then analysed to ascertain whether the qualitative findings concurred the experience and observations of stakeholders.

7.4.3 Qualitative findings

The qualitative findings were in the form of feedback from markers and comments from other stakeholders, such as panel members.

7.4.3.1 Feedback from markers

The following comments were received by email from markers of the revised scale, Model 1. These comments have been organised according to recurring themes or topics as expressed by markers, followed by the researcher’s overview of these comments (Section 7.4.3.2.).

(a) User friendliness

- “The grid is very user-friendly and allows for quick, holistic marking. This is essentially because it is all on one A4 page and is read in one direction, i.e. vertically (Top down). Also, the descriptors are very concise”.

(b) Weighting of scale

- “Am finding the grid very user friendly but have a feeling it might be favouring the weaker candidates where content and insight is thin but language pedestrian to fair”.
- “I have a ‘gut’ feel that the resulting marks are inflated when dealing particularly with weaker scripts. I am not sure if this in part is a reflection of the language component being equally weighted with content (i.e. 25/25)”
- “So, theoretically, the student can get a (borderline) fail mark for content and a good mark for language or style and pass the literature assignment even though they did not grasp the content of the literature? A moderator would have to decide which is more important in a literature assignment – use of language or understanding”.

(c) Descriptors

- “Perhaps one needs to review the descriptors under “Average” and “At Risk”. I am finding it very difficult to place a script in the “Fail” column because, whilst the language is far from the standard one would expect of academic writing, it is certainly not “barely intelligible”, as one can decipher meaning. However, placing it in the next category up immediately puts the language on a rating of 12 out of 25 which is essentially a pass”.
- “By ‘mark globally’, this marker assumed that to mean: read through all the answers and assess the whole assignment based on an overall impression guided by the rubric/criteria in the grid (as discussed at the briefing)”.
- “The headings across the top of the grid still imply comparative marking (Excellent, above Average, Average etc.).... Recommend – replace the comparative headings with the Summary descriptors at the bottom. This helps markers to place an assignment in a category – then use each criterion to establish a mark within that category”.

(d) Plagiarism

- “Is content plagiarised or is language plagiarised? ... plagiarism involves taking others’ ideas (content?) and passing them off as your own (i.e. not acknowledging the source of the idea). If it is plagiarised does it get 0? Or are there degrees of plagiarism?”

(e) Implications of comments

These comments emphasised that, while the layout was “user-friendly”, the new design did not address some of the concerns raised by markers in the previous stage (using the existing scale). Chief among these were the weighting of content in relation to language

use. Other issues were those relating to partial plagiarism and the perceived subjectivity of some of the terms. Plagiarism was of particular concern in the teaching context, as the practice has a negative effect on validity, and questions the ODL model of assignments written at home (or in other venues off-campus) and then submitted for evaluation. This is particularly concerning in the case of summative assessments completed at home. This is not the case in ENG1501, as fortunately the final examination for this module takes place at designated venues.

7.4.4 Final questionnaires – feedback on Model 1.

Feedback on Model 1 in the form of responses to the questionnaires sent to markers (Appendices H and I respectively) was summarised in Table 7.9.

Table 7.9: Summary of responses to questionnaire: Model 1

<p>Question 1 Do you think that the revised scale assesses the construct of the module better than the existing scale?</p>
<p>Results The majority felt that the module was partially better assessed. Yes: 2 Partially: 3</p>
<p>Comments a. “Assessing short answers globally on the basis of a grid (normally used in assessing compositions) instead of the allocation of marks to each question introduces a considerable degree of subjectivity”.</p>
<p>Question 2 In your opinion, is the distinction between the band levels clearer than in the existing scale?</p>
<p>Results Yes: 4 No: 0 Sometimes: 1</p>
<p>Comments a. “In the current grid, the criteria used in the band levels (vocabulary, language usage and mechanics) are too language oriented to assess appreciation of a literary piece – they assess the learner’s use of language more than their appreciation of the literature”</p>
<p>Question 3 Do you think that the increased number of levels is sufficient?</p>

<p>Results</p> <p>All agreed that there were sufficient levels</p> <p>Comments</p> <p>a. “The criteria under ‘Content’ assess the essential aspects of responding to a literary piece and the criteria under ‘Language’ assess the essential aspects of expressing one’s response fluently”.</p>
<p>Question 4</p> <p>Do you believe that the weighting of marks between organisation/content and language use on the revised grid (Model 1) produces a fairer score than the existing scale?</p>
<p>Results</p> <p>No: 5</p> <p>All respondents felt weighting was not fairer.</p>
<p>Comments</p> <p>a. “Theoretically, the student can get a relatively ‘good’ fail mark for content and a good mark for language/style and pass the literature assignment, even though they did not grasp the content of the literature”.</p> <p>b. “The problem of ‘over-scoring’ as a result of this weighting remains”.</p>
<p>Researcher’s comment</p> <p>As was the case with the existing scale, respondents believed that content should carry a greater weighting than language for this particular course.</p>
<p>Question 5</p> <p>In your opinion, are the criteria clearer in comparison to the existing scale?</p>
<p>Results</p> <p>Yes: 4</p> <p>Unsure: 1</p> <p>The majority felt that the criteria were clearer</p>
<p>Comments</p> <p>Yes-</p> <p>a. “They define the different aspects of Content and Language use at different levels for each classification”.</p> <p>b. “The grid is on one page and much easier to use. It should make it easier for the students to follow as well”.</p> <p>Unsure:</p> <p>a. “The criteria seem clear, but the weighting might give the impression that the student can pass by virtue of good language usage and not as a result of understanding the text. This should be made clearer”.</p>
<p>Question 6</p> <p>Are there any features of the scale that you think are open to misinterpretation or subjectivity?</p>
<p>Results</p> <p>Yes: 3</p> <p>No: 2</p>

<p>Comments</p> <p>Yes</p> <p>a. “The classification of Average would need some sort of benchmark example to be provided to the marker”</p> <p>No</p> <p>a. “This is an improvement on the current scale”.</p>
<p>Question 7</p> <p>Does the revised scale (Model 1) take the multicultural and multilingual distance learning target market into account?</p>
<p>Results</p> <p>Yes: 4 No: 1</p> <p>Comments</p> <p>a. “Yes, if correctly explained and used for formative testing”</p> <p>b. No: The question has two aspects: a) multicultural and multilingual b) distance. It is debatable whether distance would be a factor affecting the quality of assignments. The multicultural and multilingual aspect cannot be accounted for by a scale <i>per se</i>.</p>
<p>Question 8</p> <p>If you answered Number 7 in the negative, how can the scale be amended to reflect the distance learning context adequately?</p>
<p>Comments</p> <p>a. “The distance is not the issue. The intended outcomes and standards of the course for multicultural and multilingual users would need to be determined and then an appropriate scale adopted that would measure what it is intended to measure”.</p>
<p>Question 9</p> <p>If you could make one change to the revised scale (Model 1), what would that change be?</p>
<p>Comments</p> <p>a. “Make it two-dimensional like Model 2”</p>
<p>Question 10</p> <p>What is the main feature that you would like the revised scale (Model 1) to retain?</p>
<p>Comments</p> <p>a. “The layout (all on one page)”.</p> <p>b. “The increased number of levels in the criteria and definitions”.</p>
<p>Other comments</p> <p>a. “The problem of assessing short discrete questions globally using a grid... remains”.</p> <p>b. “The grid approach is better suited to assessing a single, integral composition. The assessor must decide in advance whether the assessment of the intended learning outcomes of the assignment would be more valid in the form of an essay or short questions. If a scaffolded essay is appropriate, taking into account the profile of the learners, the guiding questions must be designed accordingly”.</p>

7.4.5 Issues raised by the answers to the questionnaire (Model 1)

In general, markers felt that the grid was an improvement on the existing scale. There were sufficient levels and the criteria were clear. The main objection was to the weighting of content/organisation and language use. This was seen as the greatest weakness of the scale. Markers tended to favour a more integrated approach as reflected in a two-dimensional grid. Furthermore, comments such as: “In the past these aspects were accommodated by the nature and *a priori* standards set for the course being for 1st/2nd/3rd Language users” highlighted the nature of the school background of the target group, and the concomitant challenges posed by the complex teaching environment. While these problems cannot be solved by a rating scale, the target group can be assisted by a scale that can facilitate formative assessment.

The panel decided to retain the design of the grid, but to reposition certain features and alter some of the terms and wording. The summary at the end was repositioned to the top of each level to be more easily accessible to markers, facilitating quicker and easier assessment. The grid would thus have three “layers”, one being the main descriptor (e.g. “exceptional”, “good” etc.) and the others being the summary for the marker and the more detailed descriptors that could be used for formative assessment. The criteria dealing with plagiarism were retained in the content and language sections, the latter to remind markers to penalise students in both categories. The term “barely intelligible” was omitted. The final version is shown in Table 7.10 below.

Table 7.10: Model 1 grid: Final version

Classification →	Exceptional (Distinction)	Excellent (Distinction)	Good to Above Average	Average	Borderline FAIL	Seriously at risk: Fail
Mark →	25-22	21-19	18-15	14-12	11-9	8-0
Content/organisation. Criteria a. Insight: To what extent does the answer show maturity, understanding and originality? b. Awareness of stylistic/ technical features: Are these accurately demonstrated? c. Substantiation: Is the answer supported by appropriate reference to the text? d. Relevance: To what extent has all relevant information been included? Has the question been fully answered? e. Coherence – Does the answer flow together? NB MARK GLOBALLY	SUMMARY Exceptional insight and organisation. a. Original, sensitive, mature interpretation and insight. b. Exceptional, original and sensitive c. Unfailingly well-supported, shows depth and insight d. Extremely relevant, well chosen, valid ideas, all points fully covered e. Shows exceptional focus, cohesion, seamless organisation	SUMMARY: Mature, original, comprehensive. a. Thorough, incisive, original b. Excellent, good examples. c. Extremely well supported with apt examples. d. Extremely relevant, all points covered. e. Exceptionally well structured, focused, coherent.	SUMMARY: Sufficient understanding well organised, comprehensive relevant a. Good grasp of issues, some originality b. Well-demonstrated appreciation. c. Generally well substantiated. d. Mostly relevant, most issues addressed. e. Well organised, and coherent.	SUMMARY: Adequate understanding lacks originality and depth. a. Adequate, lacks depth, little originality b. Features occasionally discussed, usually correct c. Some substantiation, but mainly just thoughts on the question. d. Fairly relevant, point sometimes missed. e. Loosely organised but still coherent.	SUMMARY, Disconnected, largely irrelevant. a. Insight inadequate, little understanding of issues b. Features seldom discussed, shows lack of knowledge c. Not enough substance or relevance, insufficient support for ideas d. Many statements lack relevance e. Ideas confused or disconnected, little logical sequencing or development.	SUMMARY: serious errors, irrelevant, confused, plagiarised. a. Serious errors of understanding, extremely little evidence of knowledge of text. b. Features ignored. c. Unsubstantiated. d. Irrelevant 'misses the point'. e. Incoherent, disjointed. Plagiarised OR Not enough to evaluate.
Sub-total 25 marks						
Criteria Language and style a. Vocabulary- is there a sufficient range of vocabulary and effective word choice? b. Register and tone- appropriate for academic writing or too informal or pretentious? c. Language errors – are errors negligible or are they frequent, <u>impeding meaning?</u>	SUMMARY: Clear, fluent, articulate, sophisticated a. Excellent range and word choice. b. Very appropriate. Clear, formal c. Negligible or no language errors:	SUMMARY: Clear, fluent, articulate a. Good range of vocabulary; very appropriate word choice b. Appropriate register, competently used. c. Occasional language errors but <u>meaning not impeded or confused</u>	SUMMARY: Clear despite errors a. Word choice and range generally sufficient b. Generally appropriate. c. Some language errors but <u>meaning not impeded.</u>	SUMMARY: Adequate but pedestrian (dull) a. Small but adequate range, some errors of word choice. b. Occasional lapses c. Frequent language errors but <u>meaning seldom impeded.</u>	SUMMARY Meaning seriously impeded A. Very small range frequent errors forward choice. b. Inappropriate (too informal or too verbose). c. Frequent and serious errors, <u>meaning confused or obscured</u>	SUMMARY: Meaning severely impeded due to frequent and fundamental errors a – c Serious and distracting errors, frequently barely intelligible. OR Not enough to evaluate. OR Plagiarised

7.5 CONSTRUCTION OF THE TWO-DIMENSIONAL GRID

It was suggested that a two-dimensional scale be designed, constructed and tested as an alternative to both the existing scale and Model 1. This would potentially reduce the content versus language use problem while retaining the global marking policy. It could also be user-friendly and obviate the problem of scoring content and language use separately.

The construction process of Model 2 followed the same steps as those followed for the construction of the Model 1 grid. It also eliminated terms that had been considered vague and subjective. The number of levels remained the same as that of Model 1 and, as in the case of Model 1, criteria aligned as far as possible with the descriptors.

The most significant difference between the two alternative grids is the design of the grid and the relationship that this indicates between content/organisation and language use. This relationship is more integrated in Model 2 than in Model 1, and was an attempt to solve the possibility, raised by stakeholders, that a student could obtain poor marks for content but still pass as a result of a good language mark. The Model 2 grid makes it possible for a student who understands and appreciates the text, but lacks proficiency in language, to obtain a fair mark, while still being penalised for language use. This would be fair, particularly in the case of the large number of ESL students, many of whom are not as proficient in language skills as their HL counterparts.

It was decided to construct the grid with “content/organisation” at the top and “language use” down the side of the grid. This emphasises the importance of content/organisation while, at the same time, demonstrating the overall coherence of the scale. Students who score poorly for either content or language cannot be placed at the highest levels, thus obviating the danger of inflated scores. Once again, a summary was inserted at the top of the scale for content/organisation, to facilitate easier and more accurate marking. The descriptors for language were expressed as concisely as possible for the same reason, although it was believed that this was clear enough to contribute to formative assessment.

Table 7.11: Model 2 grid

Proposed Marking Grid ENG1501 Literary Assignments Version 1

Content and organisation NB Mark globally	A. Outstanding HIGH DISTINCTION	B. Excellent to very good DISTINCTION	C. Good to fair	D. Average	E. Fail AT RISK	F. Fail SERIOUSLY AT RISK
	SUMMARY	SUMMARY	SUMMARY	SUMMARY	SUMMARY	SUMMARY
<div> <div></div> <div>Language and Style NB Mark globally</div> </div>	Mature, original, comprehensive, very logical, exceptional insight and organisation.	Excellent understanding and organisation, comprehensive and relevant	Sufficient understanding, well organised, comprehensive and relevant	Adequate understanding, lacks originality and depth, misses some important points	Misses the point, disconnected, largely irrelevant, some evidence of plagiarism	Serious errors, irrelevant, confused. Largely plagiarised; not enough to evaluate.
	<p>a. Insight; Original, sensitive, mature interpretation and insight.</p> <p>b. Awareness of stylistic/technical features; Exceptional, original and sensitive</p> <p>c. Unfailing well-supported, shows depth and insight</p> <p>d. Extremely relevant, thought-provoking</p> <p>e. Structure shows exceptional focus, cohesion, seamless organisation</p>	<p>a. Insight: Thorough, incisive, original</p> <p>b. Awareness of stylistic/technical features: Excellent, good examples.</p> <p>c. Extremely well supported with apt examples.</p> <p>d. Extremely relevant, all points covered.</p> <p>e. Extremely well structured, focused, coherent.</p>	<p>a. Insight: Good grasp of issues, some originality.</p> <p>b. Awareness of stylistic/technical features: Well-demonstrated appreciation.</p> <p>c. Well substantiated.</p> <p>d. Mostly relevant, most issues addressed, points of question covered.</p> <p>e. Well organised.</p>	<p>a. Insight: Adequate, lacks depth, little originality</p> <p>b. Awareness of stylistic/technical features: Usually correct, but insufficiently discussed</p> <p>c. Some substantiation, but mainly just thoughts on the question.</p> <p>d. Fairly relevant, point sometimes missed.</p> <p>e. Loosely organised but still coherent.</p>	<p>a. Insight inadequate, little understanding of issues.</p> <p>b. Awareness of stylistic/technical features: Features seldom/inadequately discussed.</p> <p>c. Not enough substance or relevance, insufficient support for ideas</p> <p>d. Many irrelevant statements</p> <p>e. Ideas confused or disconnected, not enough logical sequencing, little signposting</p> <p>f. Evidence of plagiarism (whole paragraphs)</p>	<p>a. Insight: Serious errors.</p> <p>b. Awareness of stylistic/technical features: Features ignored.</p> <p>c. Mostly unsubstantiated.</p> <p>d. Irrelevant 'misses the point'.</p> <p>e. Incoherent., disjointed.</p> <p>f. Largely plagiarised</p> <p>g. not enough to evaluate.</p>
1. Outstanding. Vocabulary: sophisticated; Correct formal register effectively used; Very few language problems; <u>Meaning not impeded.</u>	A1 100%-85%	B1 84% – 75%	C1 74%-70%	D1 69%-65%		
2. Excellent to very good. Very apt vocabulary; Correct register; occasional language errors but <u>meaning not impeded or confused.</u>	A2 84% – 75%	B2 74%-70%	C2 69%-65%	D2 64%-60%		
3. Good Satisfactory vocabulary. Some errors of word choice and register but <u>meaning seldom obscured.</u>		B3 69%-65%	C3 64%-60%	D3 59%-56%		
4. Adequate Adequate of vocabulary; Frequent problems with register, language, word choice, sentence structure and mechanics; <u>Meaning sometimes obscured or confused</u>			C4 59%-56%	D4 55%-50%	E4 49%-40%	F4 39%-30%
5. FAIL: AT RISK Little knowledge of English vocabulary; poor register; <u>Numerous language problems that seriously impede communication</u> ; Not enough to evaluate				D5 49%-40%	E5 39%-30%	E6 29%-25%
6. FAIL: SERIOUSLY AT RISK Numerous problems (register, word choice, sentence structure, and mechanics) <u>that seriously impede communication</u> ; Not enough to evaluate OR plagiarised				D6 39%-30%	E6 29% -25%	F6 24%-0%

7.6 TESTING THE TWO-DIMENSIONAL GRID

The two-dimensional grid was tested both quantitatively and qualitatively, following the same procedure used to test the existing scale and Model 1.

7.6 1 Marking

Marking was carried out as for Model 1, by the same five markers, who marked 60 scripts each. The same group was used, as it was ascertained that an interval of two months had elapsed. The marking took place over four weeks. The data were then quantitatively analysed in the same way as in previous calibrations, and markers were asked to comment as for Model 1.

7.6. 2 Quantitative testing

The results of the Model 2 quantitative testing follow.

As pointed out in the discussion in Section 5.8.2.2, raters were clustered closely around the 0 logit mark as shown in Figure 7.2. This indicated good consistency between the markers. On the other hand, student ability, as seen in the “Person” column, indicated a wide range, from 3 to -3 logits. This could be a reflection on the test difficulty in relation to the test-takers and not necessarily an indication of misfitting features of the rating scale, since the scale should indicate criteria pertinent to the module outcomes for the assignment. It could be speculated that the increased detail of the scale’s criteria, as well as the changed content/language relationship, gave rise to a wider range of “person” ability. Thus, this range does not imply a misfit of the features of the scale, but reflects the difficulty level of the assignment as experienced by the target group. This had been partially masked by the relatively high marks awarded to language by the existing scale and Model 1. Furthermore, the penalties for plagiarism were operational in this scale and this could have had an impact on the lower levels of the map.

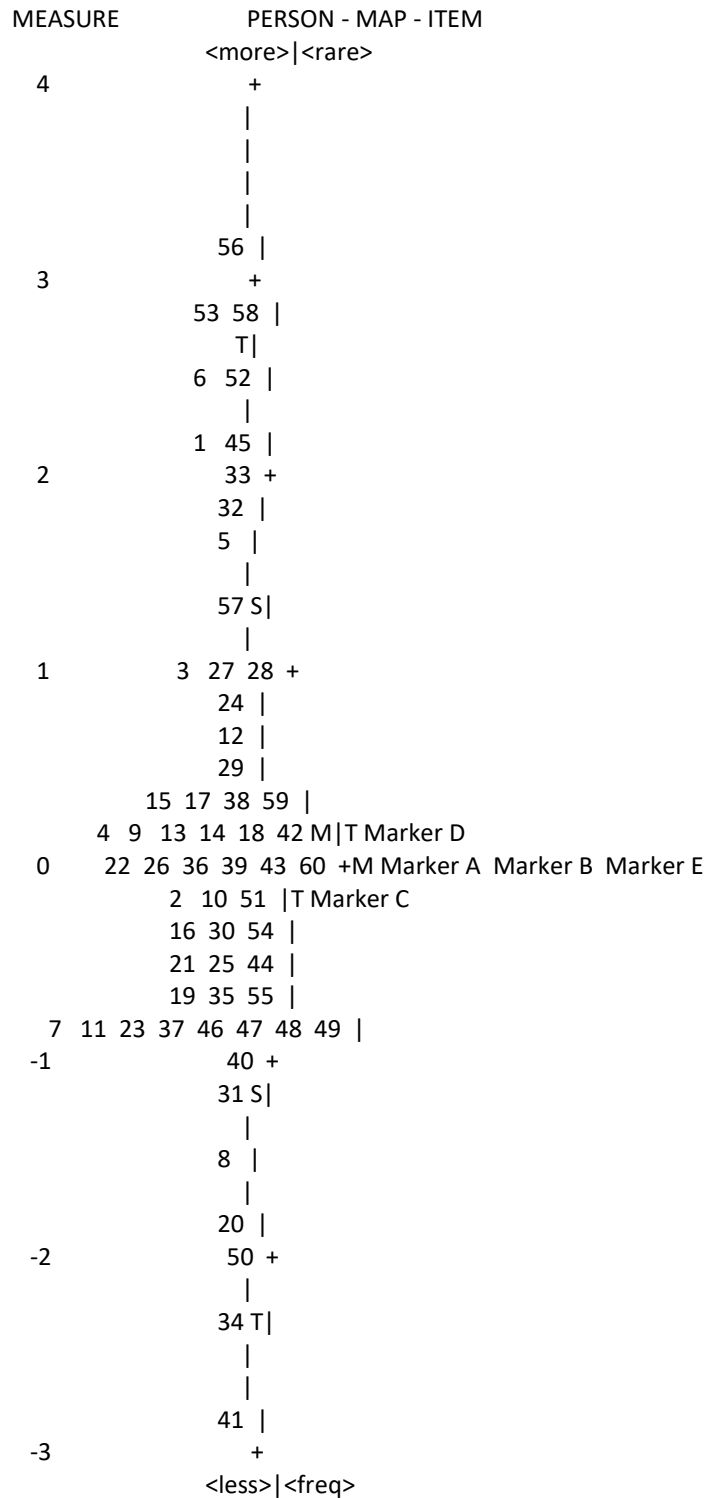


Figure 7.2: Model 2: 5 markers, 60 scripts

There were more outliers in this result than in Figure 7.2, although scripts 34 and 41 were again placed under the -1 mark. This suggests consistency between the models, but also that the Model 2 scale discriminates more finely than Model 1.

The reliability of the instrument was investigated then, in the same way as the existing scale and Model 1.

Table 7.12: Summary of reliability of Model 2

Model	Script numbers	Raters	Script Reliability	Cronbach Alpha	Rater reliability
Model 2	Scripts1 - 60	A-E	.99	.99	.75

As pointed out in Chapter 5 (5.8.2.2), where this validation was discussed as an example, the reliability of Model 2 was high in most instances as shown in Table 7.12. The rater reliability rate of 0.75 was within acceptable range, and was the result of one marker having a low mean score of 40.75 per script. In practice, this could be adjusted during marker training sessions or by discussion with the moderator or module co-ordinator. The reliability of Models 1 and 2 has been summarised in the in Table 7.13.

Table 7.13: Comparative summary of the reliability of Models 1 and 2

Model	Scripts	Raters	Person Reliability	Cronbach Alpha	“Test” Reliability	Item reliability
Model 1	Scripts1-60	A-E	.97	.97	.97	.95
Model 2	Scripts1-60	A-E	.99	.99	.99	.75

7.6.3 Qualitative findings: Markers’ comments (Model 2)

The following comments represent a selection of observations (quoted verbatim) from the markers:

7.6.3.1 Weighting

- “My overall gut feel in placing the sample scripts on this grid is that it produces a fair reflection that is a reliable assessment if holistic marking is done rather than assessing each of the six questions on individual merit. I felt this grid is effective in rating the language component as the grid does pull the script down into risky/fail when language is weak even though content suggests sufficient understanding to warrant a borderline pass”.
- “Working with the category descriptors for content, I often found relevant points under 2 categories rather than every descriptor under a category being applicable to the script (understandably as no script is going to fit neatly into a particular box unless an all-round outstanding ‘A’ or complete ‘fail’, especially with holistic/ impression marking). However, the 2 categories in these samples were

always consecutive, so this just helped to place the script finally at either the bottom or top end of the range for the selected category”.

- “I was far more comfortable using this grid to assess this assignment”.
- “In using the grid to rate literary or creative questions, we were always advised to rate the content of the assignment first and then rate language/style on that row under the appropriate column. With the marks distributed on the grid as you have them, this meant that a marker could not award an excellent language mark for somebody who showed no understanding of the content and vice versa”.
- “I find this grid user-friendly in assessing a piece holistically. I think the balance between content and language is working effectively on this grid”.

7.6.3.2 Descriptors

“Overall... I think the categories are working – I just tend to follow the detailed descriptors rather than looking at the descriptor title”.

7.6.3.3 Overview of markers’ comments.

The comments received from markers were extremely positive, particularly in connection with the weighting of marks e.g. “... this meant that a marker could not award an excellent language mark for somebody who showed no understanding of the content and vice versa” and “I find this grid user-friendly in assessing a piece holistically. I think the balance between content and language is working effectively on this grid”. The descriptors and categories also met with approval, although some were guided by the categories and others by “the detailed descriptors”. This implies that the layered design of the grid makes allowance for individual marking styles, a characteristic that would have a positive impact on validity.

7.7 Summary of results of questionnaire – Model 2

Feedback in the form of responses to the questionnaire on Model 2 that was sent to markers has been summarised in Table 7.14.

Table 7.14 Summary of results of questionnaire (Model 2).

<p>Question 1</p> <p>Do you think that the two-dimensional scale assesses the construct of the module better than the existing scale?</p>
<p>Results</p> <p>Yes: 4</p>

Partially: 1
<p>Comments</p> <p>Yes</p> <p>a. “When I marked according to the previous scale(s), I found it very difficult to decide how to categorise the script because the prompts provided by the descriptors were at times too vague and at times too prescriptive. This left me doubtful about whether or not I had adjudicated fairly in comparison with other markers and other scripts. This meant that I spent a great deal of time going back over scripts to check whether, after reading other answers, I stood by my original score”.</p> <p>Partially:</p> <p>a. “Assessing short answers globally on the basis of a grid (normally used in assessing compositions) instead of the allocation of marks to each question introduces a considerable degree of subjectivity”.</p>
<p>Question 2</p> <p>In your opinion, is the distinction between the band levels clearer than in the existing scale?</p>
<p>Results</p> <p>All respondents agreed that the band levels were clearer</p> <p>a. “In the current grid, the criteria used in the band levels (vocabulary, language usage and mechanics) are too language/mechanics oriented to assess appreciation of a literary piece – they assess the learner’s use of language more than their appreciation of the literature. The two-dimensional grid is clearer and fairer. The gradation of terminology is clearer. Thus, when one considers a term such as ‘insight’, one can grade the answer in terms of insight more easily. The same is true of terms such as ‘substantiation’ and ‘organisation’. One is not left trying to compare apples with pears”.</p>
<p>Question 3</p> <p>Do you think that the increased number of levels is sufficient?</p>
<p>Results</p> <p>All agreed that there were sufficient levels</p>
<p>Comments</p> <p>a. “The criteria under content and language assess the essential aspects of responding to a literary piece and the essential aspects of expressing one’s response fluently”.</p> <p>b. “A good number of levels is particularly important when it comes to assessing the work of the less able students. It is easy to distinguish ‘brilliant’ from ‘good’, but it is less simple to decide whether the student is really not coping. The additional descriptors help one to determine this for oneself”.</p>
<p>Question 4</p> <p>Do you believe that the weighting of marks between organisation/content and language use on the two-dimensional grid (Model 2) produces a fairer score than the existing scale?</p>
<p>Results</p> <p>All respondents felt weighting was fairer.</p>

<p>Comments</p> <p>a. “With the marks distributed as they are on the grid in this model, a marker could not award an excellent language mark for somebody who showed no understanding of the content and vice versa (except in exceptional cases with input from a moderator)”.</p> <p>b. “Yes, this is essential as clarity of thought is essential to a good answer. When the weighting is skewed one tends to fiddle the final score to allow for inadequate organisation or language use – and this is when one tends to show a bias or to become unsystematic”.</p>
<p>Question 5</p> <p>In your opinion, are the criteria clearer in comparison with the current scale?</p>
<p>Results</p> <p>All respondents agreed that the criteria were more clearly defined in accordance with the construct.</p>
<p>Comments</p> <p>a. “They define the different aspects of Content and Language use at different levels for each classification”.</p>
<p>Question 6</p> <p>Are there any features of the scale that you think are open to misinterpretation or subjectivity?</p>
<p>Results</p> <p>Yes: 2</p> <p>No: 3</p>
<p>Comments</p> <p>No</p> <p>a. “The rubric defining each level within Content/Language use is clear”.</p> <p>b. “This is an improvement on the current scale and on Model 1”.</p>
<p>Yes</p> <p>The classification of Average would need some sort of benchmark example to be provided to the marker”.</p>
<p>Question 7</p> <p>Does the revised scale (Model 1) take the multicultural and multilingual distance learning target market into account?</p>
<p>Results</p> <p>Yes: 3</p> <p>No: 2</p>
<p>Comments</p> <p>Yes</p> <p>a. “Yes, if correctly explained and used for formative testing”.</p> <p>b. “Yes, if this scale only awarded good scores to students fluent in Home Language English, one might be able to criticise it. However, it throws up good scores for those who are able to discern meaning in the work, even though they are struggling to communicate their thoughts in good English”.</p> <p>No</p>

a. "It is debatable whether distance would be a factor affecting the quality of assignments. The multicultural and multilingual aspect cannot be accounted for by a scale <i>per se</i> ".
Question 8 If you answered Number 7 in the negative, how can the scale be amended to reflect the distance learning context adequately?
Comments "The distance is not the issue. The intended outcomes and standards of the course for multicultural and multilingual users would need to be determined and then an appropriate scale adopted that would measure what it is intended to measure". "Perhaps some neutral (if possible!) graphic aids. A glossary to explain potentially difficult terms".
Question 9 If you could make one change to the revised scale (Model 2), what would that change be?
Comments a. "Change the level labels such as 'average' and 'fair'". b. "A non-verbal symbol so one could picture the essay as a well-known item".
Question 10 What is the main feature that you would like the two-dimensional scale (Model 2) to retain?
Comments a. "The increased number of levels in the criteria and definitions". b. "The two-dimensional matrix format that determines the most probable range of mark for each cell of the grid". c. "The number of graded descriptors"
Other comments a. "The problem of assessing short discrete questions globally using a grid ... remains. The grid approach is better suited to assessing a single, integral composition. The assessor must decide in advance whether the assessment of the intended learning outcomes of the assignment would be more valid in the form of an essay or short questions. If a scaffolded essay is appropriate, taking into account the profile of the learners, the guiding questions must be designed accordingly". (Researcher's comment: This comment was also made by the marker in connection with Model 1).

7.7.1 Issues arising from responses to the questionnaire

As a result of the feedback from the responses to the questionnaire, the following observations could be made.

7.7.1.1 Subjective or ‘inaccurate’ descriptor headings

There was some concern about the terms, “average” and “below average”. One panel member noted that “average is technically inaccurate” in the context of the marking grid, because, statistically, average reflects a mean, not an “adequate” score. The Model 2 grid was subsequently amended, changing “average” to “adequate”.

7.7.1.2 The distance, multicultural and multilingual contexts

It is this contextual issue that makes the current research unique. It became clear to the researcher that there are two separate aspects involved in this complex context. These are the multicultural and multilingual diversity of the target group (Section 5.6.1. – 5.6.3), and the challenges of distance learning as manifested in issues of geographical distance with the concomitant communication problems. It became obvious that a valid, clear and accessible rating scale could mitigate the difficulties caused by these extremely challenging factors, but could by no means solve them. This would require a revision of purpose and outcomes of the entire module (which was originally designed for predominantly L1 speakers). As one of the respondents to the questionnaire regarding Model 2 (Question 8) stated:

The distance is not the issue. The intended outcomes and standards of the course for multicultural and multilingual users would need to be changed and then an appropriate scale adopted that would measure what it is intended to measure.

While the researcher does not believe that the distance learning concept has no impact on students’ scores, it is conceded that the multicultural and multilingual context has a major effect on student achievement. The researcher is of the opinion that this is exacerbated by the element of distance, which prevents adequate communication between stakeholders. In the present situation, a valid rating scale, which is accessible to students and markers alike, which can promote ease of marking while addressing the given outcomes of the module adequately, and which can provide formative guidance to the students, is the aim of the present research. However, the outcomes could be re-worded to specify or clarify grandiose-sounding (but arguably hollow) terms such as: “The dimensions of artistry and contrivance” to include descriptors given on the rating

scales, such as “appreciation of technical features”, “understanding of the use of imagery” etc. Ultimately, the structure of the module might have to be altered, or (preferably) another module could be introduced, dealing with foundations of literary appreciation to cater for L2 students. This could possibly take more L2 language groups and educational backgrounds into consideration.

7.7.1.3 Plagiarism

The Model 2 grid, in particular, makes provision for dealing with plagiarism, and instructs markers to penalise partial lifting of sources without acknowledgement, as well to penalise heavily in the case of the whole script or large portions of it being plagiarised. Plagiarism must be addressed as a matter of urgency, as it has a serious effect on validity. While it was believed that plagiarism should be heavily penalised, the panel also conceded that some marks should be awarded to students who did not plagiarise the entire answer. Various solutions were discussed, and various degrees of severity were debated. Ultimately, the decision was made to score plagiarised and partially plagiarised scripts in the lowest two levels, as shown on the grid.

7.8 ACTIONS AND SUGGESTIONS ARISING FROM FEEDBACK MARKERS AND PANEL MEMBERS

Although the issue of assessing short questions using the holistic current grid was not raised as a problem by the Unisa markers when evaluating the existing scale, feedback from some markers of the Model 1 and Model 2 experimental grids included reservations about using the grid to assess short questions. As a result of this feedback, the panel took the following actions, in addition to the alterations on the grids as described.

7.8.1 Assessing the essay-type question

The problem of assessing a contextual-type poetry assignment by means of a global rating scale was mentioned by stakeholders. To address this concern, the researcher had obtained a very small number of essays ($n = 4$), as opposed to the shorter answer format

of the poetry assignment, written in the same semester as the target assignments. The students had consented to the use of these scripts and, in the case of the scripts marked by Unisa markers, had sent the marked script to the researcher for this purpose. The other scripts were written for a mini-examination set by the researcher in the course of tutorial classes at the Parow Regional Centre of Unisa.

Since this exercise was based on a very small sample ($n=4$), it was not possible to calibrate the marks statistically, but the following results show consistency among the markers (in this case, the panel members). Markers are indicated by M1, M2 etc., with Script 1 M1 of Script 1 indicating the original Unisa marker who marked according to the existing scale. Scripts 2 – 4 in the M1 column were marked by the researcher also using the existing scale. The other scripts were marked by the panel members using the Model 2 rating scale.

Table 7.15: Results of assessing essay-type answers

Script	M1	M2	M 3	M4	M 5
1	90	83	80	82	83
2	50	53	54	55	50
3	32	32	30	30	32
4	66	60	64	64	64

The only large discrepancy was found in the result of Script 1, which the original marker had awarded a much higher mark than that given by the other markers¹³. It was not possible to ascertain a reason for the original high mark as other marks given by this marker were not available. However, this exercise did give some indication of the reliability of the two-dimensional grid in assigning marks to both the short question (poetry) assignments and the essay-type assignments.

7.8.2 The use of graphics for formative assessment






While the panel agreed that the use of graphics could contribute to formative assessment, choosing suitable graphics could present a problem in the diverse and multicultural target group. Two examples, shown in Figures 7.3 and 7.4 below, were

¹³ It was noted that a variation of 6% is not that large, while greater than one would wish. Analysis can deal with more than 2 parameters, and marks can be adjusted during the moderation process

discussed, but it was not certain whether these images were sufficiently culturally neutral.

Table 7.16: Grid for assessing Grade 12 literary essays (First Language, Higher Grade

MARKING GRID

English First Language (HG): Paper 1	EXPRESSION O →	A. POWERFUL Style ¹ is deliberately employed to enhance meaning/outstanding use of linguistic options/ errors ² of no consequence 	B. PRECISE A fluent and competent writer/competent use of linguistic options/errors of minor consequence 	C. ORDINARY Fair use of linguistic options/errors distracting 	D. HALTING Limited use of linguistic options/errors serious. 	E. DIFFICULT TO UNDERSTAND Hardly any use of linguistic options/errors are very serious 
↓ CONTENT ↓						
1 OUTSTANDING Original/creative/ comprehensive/fuck/nature		54-60 36-40 A+	48-53 32-35 A	42-47 28-31 B		
2 GOOD TO VERY GOOD Interesting/solid/convincing/ effective/imaginative		48-53 32-35 A	42-47 28-31 B	36-41 24-27 C	30-35 20-23 D	
3 RELEVANT BUT UNINSPIRING Unoriginal/predictable/ occasional/ flashes of insight		42-47 28-31 B	36-41 24-27 C	30-35 20-23 D	24-29 18-19 E	18-23 15-15 F-FF
4 UNCONVINCING Inappropriate/contrived/ lack of logical progression/thin			30-35 20-23 D	24-29 18-19 E	18-23 12-15 F-FF	12-17 8-11 G
5 UNACCEPTABLE Contradictory/often incoherent/off the topic/most unlikely				18-23 15-15 F-FF	12-17 8-11 G	0-11 0-7 H





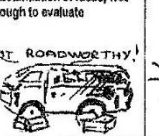

¹ Style: The chosen method of expression using a variety of linguistic options.

² Linguistic options: Syntax (variety of sentence types, lengths and structures; register; tone; appropriate paragraphing and punctuation; choice of imagery, adjectives, verbs and adverbs)

³ Spelling, case, tense and concord are some of the errors we should be considering.

Source: (Cape Department of Education, n.d.)

Table 7.17: Grid with graphics depicting mini-bus taxis

	A Outstanding	B. Excellent to very good	C Good to fair	D Average to below average	E Shaky	F Very Shaky
→	Very thorough understanding of text; Relevant, original arguments; Mature ideas; Exceptional appreciation; Awareness of technical and stylistic features of text; Succinct, well-organised; Well-supported arguments	Thorough understanding of text; Relevant arguments; Mature ideas; Good insight; Thorough understanding of technical and stylistic features; Well organised; Ideas well substantiated.	Fair understanding of text; Mostly relevant arguments; Some immature ideas; Fair insight; Adequate understanding of technical and stylistic features; Loosely organised, incomplete sequencing; Ideas not always substantiated	Superficial understanding of text; Arguments frequently irrelevant; Some immature, superficial ideas; Adequate insight and understanding of technical and stylistic features; Muddled organisation, but sometimes logical; Some substantiation of ideas	Little understanding of text; Predominately irrelevant arguments; Immature ideas; Little insight; Little understanding of technical and stylistic features; poor organisation ideas confused, little logical sequencing insufficient substantiation of ideas; Not enough to evaluate	No understanding of text; irrelevant arguments; Immature ideas; No insight; No understanding of technical and stylistic features; Poorly organised; No substantiation of ideas Not enough to evaluate
↓ style						

Source: Adapted from proposed Model 2 scale. Designer: Lindsey Lewis

The panel was of the opinion that both sets of graphics are imaginative in their attempt to describe the criteria in a non-verbal fashion. The team members who had used the grid depicting motor cars (Figure 7.3) praised it for giving a light-hearted and clear description of the levels. The panel felt that the use of graphics would address some of the challenges raised by the demographic profile of the students in a creative and original manner. Although no decision was reached to include graphics in the scale

design at present, it was decided to recommend this possibility as an area of further research. The study material could also include examples of graphics to explain the grid to the students.

7.9 CHOICE OF RATING SCALE

Both grids designed by the panel were deemed to be an improvement on the existing scale. The relative advantages and disadvantages can be summarised as follows:

- Both scales demonstrate construct validity in their alignment to the stated outcomes of the module.
- As regards scoring validity (reliability), both were statistically valid, with Model 1 showing slightly higher Item reliability.
- The weighting of content versus language use in Model 1 was shown to have a negative effect on construct validity.
- The weighting of content versus language in Model 2 provides a fairer, more coherent and more balanced relationship between these two major criteria.
- Model 2 is seen to have greater face validity, which implies that it is intuitively more appealing.

Markers and panel members were essentially unanimous in their preference for the two-dimensional Model 2 grid. Thus, it was decided, after alterations to the wording and minor adjustments to the scale layout, to recommend that the version of the two-dimensional scale as presented in Table 7.11 be adopted.

However, it was conceded that Model 1 could also be considered a viable alternative, provided markers were cautioned not to award high marks for language in the case of scripts that show little or no knowledge of, and insight into, the literary text. The panel debated at length whether it would be possible to present both scales as possible replacements for the existing scale, but were unsure whether this would be accepted in the context of this thesis, which required that one scale should be chosen.

7.10 SUGGESTED IMPLEMENTATION

The following examples demonstrate how the two-dimensional grid is used in marking. These scripts were marked for the purpose of the research, but comments made by the

marker can be adapted to suit the formative function of the feedback. In reality, more in-text remarks (for instance on language) can be given. Markers' comments are provided in italics in the examples below, and are used as examples of comments that could be given to students.

Examples 1 and 2 demonstrate the differences between two scripts that would be classified as Level 3 in the existing scale. They also demonstrate how the grid does not allow more fluent students to pass if the literary content is misinterpreted.

Table 7.18 Example 1

Example 1

1.

Contrast is created by using sympathetic words while describing the black women with loss and white women with loss is described unsympathetically *By the poet or by the doctor?* "Mandela's daughter tried to find her father through the glass. She thought they'd let her touch him." (line 24 & 25). It also takes place in line 26: "And this woman's hands are so heavy when she dusts..." which suggest sadness. In line 33 "They (the nannies) talk about everything, about home..." which suggest how they long for their homes. *How does this demonstrate contrast?*

The white woman's *sp.* loss is degraded to nothing, line 1 and 2 "...you may not suffer the death of your stillborn." She is urged to stop mourning. The death of her child is considered a "...small passing(s)" (line 14) not necessarily because it was a baby, but according to this man, *this* is insignificant compared to the loss of the black community. *Does he actually say this? Substantiate*

She is even forbidden by this man *substantiate* to not even mourn in the privacy of her own home. "Do not circle the house, pack, unpack the small clothes." (line 8 and 9).

2.

The words "...you may not suffer the death of your stillborn..." (line 1 & 2 repeated in line 13 & 14) evoke the idea of bitterness. The repentances (*repetition?*) of "Do not..." (line 6, 8 and 10) emphasises the bitterness of the man speaking these words.

It is shocking that a man *should* speak so heartlessly to a woman going through so much pain, physically and emotionally.

These expressions in stanza 1 convey *concord* a bitter sense of shock and bitterness.

In stanza 2 ideas of shocking and upsetting images is

conveyed. "...a newspaper boy in the rain..." (line 15) suggests that a "...boy..." (line 15) which is working, who is still schoolaged "...sleep...in a doorway." (line 16) which implies that he has no home and/no one to care for him. "...baby ... sent to a tired aunt ... return a stranger" (line 21-23) evokes the idea of parents unable to care for their children. "She thought they'd let her touch him." The speaker seems critical of the prisoner laws of that time. *Substantiate. Relevance?*

These images are not pleasant ~~to form in your mind.~~ *it is They are* used to suggest the idea of a man *who is* very upset(?) telling *Explain - unclear* telling a woman whose

baby died stillborn that her sorrow is not nearly as much as that of the black women. The words he used implies torture (?). *Substantiate with examples of these words* Stanza 3 implies sympathy (*from whom?*) "...woman's hands are so heavy..." (line 26) implies the sadness that overwhelms her when she "...dusts the photographs of other children..." (line 26 & 27). "They talk about everything, about home..." (line 33) This evokes a feeling of sympathy as the nannies are with children, but not their own and that they (*Who? Ambiguous*) are at a home, but not their home.

Part 2, Stanza 4 also conveys a torturing and upsetting sense (?) "Child shot running..." (line 39) and "...boy's swollen stomach full of hungry air." (line 41 and 42) is devastating *Substantiate*.

Part three, stanza 5 conveys a comforting sense with the repetition of "They will not..." (line 50, 51 and 52) (*Explain why this is 'comforting'*) followed by the cruel words of stanza 1 (*Quote*). This suggests that the black mothers who have lost a child/children will not see the white lady's loss as unimportant, but will see it as something to mourn about.

These tones enhances (*concord*) the pictures the poet give to us as readers, it makes it more emotional. *Explain*
3.

The main sound device used by the poet is repetition.

"In this country you may not..." (line 1 and 13) conveys a sense of deprivation, the same woman with the same loss in a different country will not be urged to stop mourning, but death is death even in another country, so is mourning! This man who spoke these words evoke the idea of wanting to degrade the white woman's loss against the losses of the? Repetition of the words "Do not..." (line 6, 8 and 10) emphasises that the person saying this wants to take away from the white woman that he would not take away from a black woman suffering the same loss.

Repetition of the words "They will not..." (line 50, 51 and 52) conveys a sense of assurance that they will not adopt the same cruel attitude towards someone with loss as they (the black woman) have experienced it and do not wish it upon their supposed enemy. They will comfort the white woman with loss(?) as they know what the comfort is worth in difficult times. These *This* repetitions ~~have~~ *has* a great (*vague*) effect on the poem, it forces the reader to see it and to understand the value of the repetitions.

Longwinded

4.

Two similes from the poem is *are* "...their skins like litmus..." (line 35, stanza 2) and "Their mourning rises like a wall..." (line 48, stanza 3).

The simile of line 35 ("...their skins like litmus...") is used to suggest the idea of contrast between the children(s)..." (line 34) skin colour (white) and the skin colour of their "...nannies..." (line 31) (black) who is *are* (*concord*) looking after them. The second simile implies that the grieving of black mothers who has *have* lost a child/children is *are* protected ("...wall..." line 48) from the opinions of other people. *A wall is seen as an obstacle or something difficult to surmount. Note that it is 'rising'. The black women in this poem do not seem to be 'protected' from the opinions of others or from the loss of their children. Read the poem carefully – where is the substantiation for your statement?*

These two similes emphasise the suggestion in the poem that life of any person is important and thus their deaths too. *How do the two similes emphasise this, especially the 'skins like litmus'?* The white mother was unsympathetically told "...you may not mourn small passings." (line 13 and 14). The speaker implies that black mothers who have lost child/children will not consider the death of any child more important than that of another child. ("They will not compete for the ashes of infants." (line 52).

6.

Line 53 to 58 evoke the idea of caring from people the community does not expect whilst the rest of the poem implies an idea of cruelty even from people whom the community thinks should be caring. "I think they may say to you: Come with us to the place of mothers." (line 53 and 54) opposed to "the doctor says 'It was just as well' and 'You can have another.'" (line 11 and 12) *Good* The black women who have suffered great loss invites the white woman with them and the doctor pushes the white woman away with these unsympathetic words.

"We will stroke your flat empty belly..." (line 55) is in contrast with "...boy's swollen stomach full of hungry air." (line 41 and 42). The black women will comfort the white woman who lost her baby, but they do not have the opportunity to feed their own children whose tummies is *are* swollen because of hunger.

"...let you weep with us in the dark..." (line 56) is contradicting line 15 and 16 which states "See: the newspaper boy in the rain will sleep tonight in a doorway." The boy will be sleeping and most probably weeping alone as only one person can sleep in a doorway whilst women in a similar situation as his mother (alone) will comfort a woman who has lost her child as a lot of them has *concord* lost their children, not only by death, but by returning "...a stranger." line 23. "... and arm you with one of our babies to carry home on your back" (line 57 and 58) is opposing line 37 "Baby no one carried alive..." When this white lady has another child, the people surrounding her would stop telling her to mourn the child of her that was stillborn. It is opposing (?) as it will not be her baby, but the baby of a black mother.

D4 55%

D4 Long-winded; does cover the basics, makes some good points but adds much unnecessary material.

D. Adequate.

Feedback: Relevant points (copied from grid)

SUMMARY

Adequate understanding, lacks originality and depth, misses some important points

a. Insight: Adequate, lacks depth, little originality

b. Awareness of stylistic/ technical features: Usually correct, but insufficiently discussed

c. Some substantiation, but mainly just thoughts on the question.

d. Fairly relevant, point sometimes missed.

e. Loosely organised, but still coherent

Table 7.19 Example 2

Example 2

Question 1

A woman has to deal with the shock of losing her baby in a society where death is an everyday reality, or taken as a norm. Here a poet emphasizes that a black woman (*the woman in the poem is white – read the introduction to the poem*) cannot mourn the loss of her still born child, ^^ the way in which they are abused. After losing a child (still born) she has to go back to work as nothing has happened. She has to leave her children go and look after whites' children. When she is there she felt *feels* the pain when she dusts the photographs of her employer's children. Black women suffer a lot. *They* cannot even share their pain, they have to meet at *on* pavements to share their pains or show the other maid that they feel her pain. They are able to comfort her and see her loss as a genuine catastrophe which is indeed comparable with all other tragedies happening around them. Hers is literally no small passing. Women has *have* (concord) to stand together against the abuse they are facing in South Africa.

Question 2

The tone is bitter. - the country which she lives should have such an indifferent attitude towards the woman's suffering. She is bitter she is not allowed to ^^ the loss of her own child. It compares the suffering the woman had with what South Africans experience in their everyday life. Women being abused daily, men tell them what and what not to do. *Incomplete sentence*. They are not free. They have to meet on pavements to talk about everything (*is this still the case?*). It just reminds every woman *sp*. of every abuse they come across every day. To be told what to do and what not to do. *Incomplete sentence*. They have to lift up their heads and pretend as if nothing has happened. With tears *in* her eyes she has to say everything is fine. The abuse cannot be physical always it can be emotionally, discriminatory, sexually, financially etc. Mothers have to care for their children under each of the situations and make them believe that everything is under control. They (women) have to overlook oppression. *Focus on the poem*.

Question 3

They – the poet says that the nannies who have lost ^^ understand her suffering. They will not tell her that her suffering is less important than theirs. They understand how deeply a mother is affected by the loss of a child and they will sympathise with her, even though she is white and regarded as privileged *sp*. They understand that her privileged *sp* status does not mean that her suffering is less important than theirs. She suffers the anguish of her child's death just as much as they do. The mothers who have lost their children due to politics such as when they were shot dead while stoning police. People alluded to such incidents when telling her that her suffering was as great (?) as that of others. *Question not answered (QNA)*

Question 4

The first simile *Quote* Litmus is a white-coloured paper which is used to gauge and test for acidity and alkalinity; it turns reddish in the presence of acid, and blue in the presence of alkaline. Like the paper children are white and are in contrast with their black nannies who are caring for them. The fact that there are white children at this gathering is a gauge or indication that this is not a political but rather a social gathering. The presence of white children is the indication that the black nannies are not

contravening the South African Riotous Assemblies Act of 1956 which forbade the public gathering of three or more people. *QNA*

Question 5

Mothers cannot compete on how many of them have lost their children but they can meet comfort and sympathise with each other. Their loss is not a competition it's a severe painful loss, yet they are told not to mourn for their children.

Mark E4 40% *Many questions misinterpreted.*

Bibliography

Tutorial letter 101/3/2016- ENG1501

Season come to pass- small passing page 254-255

Internet

Key- QNA= Question not answered

Copied from grid:

Misses the point, disconnected, largely irrelevant.

- a. Insight inadequate, little understanding of issues.
- b. **Awareness of stylistic/technical features;** Features seldom/inadequately discussed.
- c. Not enough substance or relevance, insufficient support for ideas
- d. Many irrelevant statements
- e. Ideas confused or disconnected, not enough logical sequencing, little signposting

The following two scripts demonstrate examples of extensive plagiarism and partial plagiarism respectively.

Table 7.20 Example 3

Example 3

Small Passing (Ingrid de kok 1951-)

1. The epigraph introduces a stark contrast between the 'small passing' and the everyday suffering of black south africans.by referring to the first section(lines 1-35)of the poem

the effect of the opening section, that "in this country you may not/mourn small passing" (line 1-2), is a powerful indictment on the state of the country and an acknowledgment of the magnitude of suffering of black women.it is possibly a genuine prohibition towards the women mourning, or simply a descriptive statement, rather than a sarcastic response to the man's worlds in the preamble. The small 'white' tragedy

which in other circumstances deserves to be mourned, may not be mourned; the speaker, by virtue of her inescapably white (read oppressive) subject position, forgoes entitlement to suffering. The images of national suffering are preceded by an instruction to “See” (line 15)-become aware of and to understand the plight of others, and to gain perspective and some measure of humility even amidst personal grief.in a country where suffering has been so unquestionably. ‘Black’ it is too easy to say that suffering knows no race or colour.as servants in the homes of white women, meet on pavements in areas where they work rather than in their homes. They take the domain of motherhood with them. The image of nannies and the allusion to the designation of space importantly recognise the power relations between white and black women that impede community and solidarity. *This paragraph has been plagiarised*

2. Tone of poem and how it contributes to its meaning.

Undermines the haughty superciliousness and utterly objectionable tone and tendencies displayed by the doctor in his arrogant, “it was just as well”(line 11)and “you can have another”(line 12).while an argument can be made that the views expressed by the doctor should be seen as nothing more than vocational professionalism from someone who has seen such loss more often than not and may, within this context, be suggesting, a pragmatic way forward it needs to be remembered that the persona has already created a specific context within which to understand the doctor’s statement, in the poem’s dedication: for a woman whose baby died stillborn, and who was told by a man to stop mourning, ’because the trials and horrors suffered by black women in the country are more significant than the loss of one white child(my emphasis).against the foregoing background such statements coming from a doctor who be more sympathetic are crass, to say the least. *Marker’s comment – where is the “emphasis”? Name the reference.*

This paragraph has been plagiarised. Quoted from MC Mashige.

[Researcher’s note: See similar wording in the next example]

3. Sound device the poet employs and its effect by referring to at least three examples from poem

the main sound device the poet employs ‘alliteration’ *Marker’s suggested comment to student: Why do you use the quotation marks?* Alliteration is also called head rhyme or initial rhyme, the repetition of the initial sounds(usually consonants)of stressed syllables in neighbouring words or at short intervals within a line or passage, usually at word beginnings, as in “ like litmus”(line 35)”boy’s swollen stomach”(line 41)”shadow and silence”(line 3).sidelight: the sound of alliteration produce a gratifying effect to the ear and can also serve as a subtle connection or emphasis of key words in the line, but should not “call attention” to themselves by strained usage. Alliteration often works with assonance and consonance to make phonetically pleasing arrangements.

Alliteration sometimes is very subtle. When we study alliteration, we are concerned with the sounds of the words, not just the letters. *Marker’s suggested comment to student: Use your own words. If you are quoting from another source, you must give the reference. Explain the effect of the examples that you mention*

4. Two simile from the poem and what the mean in the poem as a whole

1. “Like litmus” (line 35) Litmus is a paper that turns red in acid or blue in an alkali. Such a description acts as a double-edged sword that cuts both sides of the divide *Tutor’s comment – use your own words.* On the one hand the children have lotion applied on their faces to protect their skins from the very hot African sun.to assist in their skins’ protection they also put on bonnets, which provide some shade for their skins. On the other hand the description shows that the nannies are able to survive even in the face of the harsh African sun, which the children cannot survive without some form of skin protection.2. “Rises Like a wall” (line 48) on the one hand, the statement

refers to a deep sense of loss that the mothers naturally, mourn; on the other hand, it engenders a sense of hope and reflects the courage that the mothers have. Mourning in the stanza has echoes of mourning *sp.*, which marks the beginning of new day, and thus hopes for better things. *Marker's comment: Not a simile*

Marker's comment- this answer has been copied verbatim from a reference.

6. Line 53-58 they contradict the rest of the poem

In its examination of the role of motherhood, the poem represents “the place of mothers” (Line 54) as one imbued with qualities such as compassion, comfort, racial harmony and empathy, in short, a place where mothers console each other and thus form a solidarity that helps in articulating their identity as women irrespective of their racial and class origins. Here one detects tension between the old and “new” motherhood has often been used as way confining women to conservative roles as minders of the household. That the persona projects a vision in which women come together in solidarity, almost in open defiance of their conservative, socially constructed roles of domestication and racial difference, is a sharp departure from the apartheid construct of white “madam”/black “nanny” power relations. As a result her determination to see a more just society, the persona envisions a society in which white women enter a “place of mother”, stripped of the pretensions of their assumed power and superiority over their black counterparts. *Marker's suggested comment to student: See comment above. Use your own words and observations.*

Biography

Locket, C. 1996. *Feminism(s) and writing in English in south Africa*. in daymond, M. J. (ed.). *South African Feminisms: Writing, Theory and Criticism 1990-1994*. London: Garland Publishing inc, 3-26
De kok, . 1998. *Familiar Ground*. Johannesburg: Ravan Press
Daymond, Margaret (ed). 1996. *South African Feminisms: writing, theory and criticism 1990-1994*. New York London: Garland Publishing
<http://www.Google/Poetic devices/> (15 August 2016) 21 August 2016

F6	20%
-----------	------------

Marker's comment: Suggested response to student: This assignment has been largely plagiarised. Please revise the notes on plagiarism, and refer to the grid, which indicates how plagiarism is penalised. Please remember that you submitted a declaration that this was your own work. We want to hear your voice, not someone else's.

Table 7.21 Example 4

Example 4 Partial plagiarism

1. The epigraph introduces a stark contrast between the ‘small passing’ and the everyday suffering of black South Africans. By referring to the first section (lines 1-35) of the poem, explain how the poet creates this contrast.

The poem deals with the shock of losing one's baby in a society where death is an everyday reality, a woman whose baby dies stillborn, and who was told by a man to stop

mourning, “because the trials and horrors suffered daily by black women in this country are more significant than the loss of one white child.” This dedication invites racial and gender-based interpretations of the poem. Because of the history of race imposed by apartheid, the man referred to in the dedication seems to think that the pain of one woman’s loss of her child is irrelevant compared to the suffering of black children and all around her in apartheid South Africa, where death is the norm. “And this woman’s hands are so heavy when she dusts the photographs of other children they fall to the floor and break. She moves so slowly as if in a funeral rite”. The contrast would be how a mother of a white child has lost her baby and should not mourn, while the mothers of their black children “lost” their own children by having to come and work for white people and left their own children behind to look after the white children. She is feeling sad that she cannot spend time with her children. She is moving slowly like she is in a burial service as she dusts and cleans the house.

Marker’s comment: The response shows a clear understanding of the contents of the poem. Quoted sections of this response indicate that the student has researched the work of others. No in-line referencing and does not acknowledge sources, however. Much of the response has been borrowed directly from the work of others. This negates the value of the response.

2. Identify the tone of the poem, and explain how it contributes to its meaning.

The poem covers anger, fear, love, pain, and suffering and eventually hopes for the dawn of a new era, one in which there is compassion, sympathy and care. The poet is bitter that the country in which she lives should have such an indifferent attitude towards her suffering. She is bitter and sad that she is not being allowed to mourn the loss of her own child. The poet contrast these mothers with the people mentioned in stanza 1 who kept repeating: “Do not.” In doing so, she points out the difference between people who do not understand the impact of the death of the child and mothers who do sympathise with her mourning. The poet points out how easy it is for those who do not understand to give advice, telling her to pull herself together.

Marker’s comment: Effective response

3. This poem is an example of free verse, which means that it has no set structure.

However, the poet uses a number of poetic devices to create rhythm and form. Identify the main sound device the poet employs, and discuss its effect by referring to at least three examples from the poem.

“The useless wires and cords on your stomach, the nurse’s face, the walls, the afterbirth in a basin.” These are reminders of the painful process of having to face up to the loss of a child. This is a stylistic device the poet uses to indicate that the loss of a child is a painful experience to mothers, no matter what their race is, religion and even political origins and inclinations. Reading the poem one is struck by the deep sense of loss for a mother, not just a white mother. “It was just as well.” And “You can have another.” The words expressed by the doctor should be seen as nothing more than insensitive from someone who has seen such loss more often. “For a woman whose baby died stillborn, and who was told by a man to stop mourning, because the trials and horrors suffered by black women in this country are more significant than the loss of one white child.” Such statements coming from a doctor who should be understanding and be more sympathetic are insensitive to say the least.

Marker’s comment: See comments regarding Question 1. Loose commentary surrounds a ‘lifted’ response from the work of MC Mashige, which references the work of Bowen.

4. Quote two similes from the poem, explain what they mean. Also consider how these similes develop meaning in the poem as a whole.

“Their skins like litmus.” Litmus paper is used to gauge and test for acidity and alkalinity. *Marker’s comment: Once again, the student has ‘lifted’ his/her response in part.* It’s initially white in colour but turns reddish in the presence of acid and blue in the presence of an alkaline. Like the paper, the children are white, and are in contrast with their black nannies who are caring for them. The fact that there are white children at this gathering is an indication that this is not a political gathering but rather one of a social nature. The presence of the white children is an indication that the black nannies are not contravening the South African “Riotous Assemblies Act” which forbade the public gathering of three or more people. “Their mourning rises like a wall.” The statement refers to a deep sense of loss that the mothers naturally mourn; on the other hand it gives hope and reflects the courage that the mothers have. “Mourning” in the stanza has echoes of morning, which marks the beginning of a new day and hope for better things. The mothers’ mourning gets a sense that their hope “rises like a wall” a wall of hope on which they see possibilities not only for new beginnings but hope for the future.

Marker’s comment: This too has been taken directly from the work of Mashige. This source has not been listed as a reference.

5. Carefully consider lines 53-58, and explain how they contradict the rest of the poem.

The poet says that the black mothers who have lost children understand her suffering. They will not tell her that her suffering is less important than theirs. They understand how deeply a mother is affected by the loss of a child and they will sympathise with her, even though she is white and regarded as privileged. They understand that her privileged status does not mean that her suffering is less important than theirs. She suffers the pain of her child’s death as much as they do. The poem represents “The place of mothers.” As one fills with qualities such as compassion, comfort, racial harmony and empathy, in a short place where mothers console each other. Any death, even that of an ordinary baby is important and causes suffering. People must stop telling mothers thus affected that their suffering is not as great as the suffering of others. All suffering is important. *Marker’s comment: Competent response. However, given that much of the rest of the response has been plagiarised, it is not certain whether this is original or not. There is also only a vague reference to the question, which refers to how the last lines contrast with the rest of the poem.*

Mark: E4 40%. Evidence of plagiarism

List of sources

Bowen, B.E 1996. South African Feminisms: Writing, theory and Criticism. London. Pages 3-50

Moffett, H. 2013. Seasons come to pass. 3rd ed. Oxford University Press. Cape town, South Africa. Page 254

Ruthven, K.K. 1984. Feminist Literacy Studies. Cambridge: Cambridge University Press. Pages 16-32

Lewis, D. 1992. The politics of Feminism. South Africa. Pages 15-21

7.11 CONCLUSION

In this chapter, the process of developing a new scale, based on the quantitative and qualitative findings has been explained. The panel discussions were summarised and the stages of developing a new scale, namely: the design stage, the construction stage and the trial stage) were discussed in detail. Evidence includes quantitative elements, in the form of statistical analysis, as well as the qualitative features extracted from the comments of markers employed at various stages of the process. Finally, the proposed new assessment scale was presented and reasons given for this choice. The chapter concludes with a summary of the process followed and the results of this process.

Chapter 8: Conclusion and recommendations

In any measurement exercise, the measure must be considered to be valid, to assess what it claims to measure, and to be consistent by producing reliable outcomes (Mahlangu, 2016: 111).

8.1 INTRODUCTION

The aim of this study was to investigate the validity of an existing assessment scale for academic writing in response to literary texts in a distance education environment and, depending on the outcome of the investigation, to modify the existing scale in order to produce an empirically validated scale for assessing the assignments of the target group (ENG1501 at Unisa). The problem thus addressed was the validation of a rating scale appropriate to its purpose and context.

In this final chapter, the findings of this study are summarised, discussed and recommendations are made on how these findings can be implemented. The recommendations have been structured according to the primary and secondary research questions posed at the beginning of this thesis. The findings relating to each question are discussed, followed by recommendations arising from recurring themes identified during the research process. The limitations of the study have then been noted, after which suggestions for further research are made.

The research was undertaken against the background of the body of theory generated by debates on validity, the ODL context, and the theories of literature teaching, particularly in the university environment and with reference to the challenges of South African socio-economic, linguistic and cultural diversity. The research process took into account theories of validity (Chapter 3) and validation (Chapter 4), and was based on the argument-based validation approach (Shaw & Weir 2007, Kane 2017).

8.2 QUESTIONS

The primary research questions were two-fold as follows:

1. Is the existing assessment scale used for the *Foundations in English Literary Studies* (ENG1501) at Unisa valid in terms of the various aspects of validation and purposes (namely formative assessment, summative assessment, feedback) and
2. Depending on the results of the study, how can the existing scale be modified or replaced in order to produce an empirically validated scale for assessing the assignments of the target group?

These primary questions were supported by sub-questions, and they were all addressed by using a mixed, qualitative and quantitative method.

Quantitative research comprised the statistical processing of the data, while the qualitative features were characterised by stakeholder input. This stakeholder participation was in agreement with the observation of Johnson *et al.* (2015: 129) that “there is a strong rationale for research to evaluate policy impact by gathering information directly from those individuals who directly experience it”. In the current study, the observations of markers, tutors, lecturers and panel members were examined carefully in order to evaluate the present rating scale and to design alternative scales that reflected the construct more accurately, while also facilitating formative assessment, an important feature in a distance education context.

The findings relating to each research question are discussed in this chapter. Recommendations incorporating the recurring themes generated by all stages of the research then follow.

8.2.1 Sub- question 1

What do the results of the empirical research process reveal about the validity of the existing scale?

The concept of validity was discussed as part of the theoretical framework of this thesis. The debate included differing opinions on the relationship between the various types of validity, as well as their relative importance. The relationship between validity and reliability was also discussed, and it was decided, in keeping with modern interpretations of validity, to incorporate the two concepts, with reliability (or scoring validity) seen as an aspect of validity. Construct validity was considered to be the overarching type of validity, influencing the relationship between other validity types.

Theoretical models and frameworks were examined, and the argument-based approach which consists of a claim substantiated by supporting evidence (Shaw & Weir, 2007), was adopted as a framework for the research process. Factors that affect scores were also taken into account in this study, such as assessor characteristics and training, and inter-rater discrepancies. This process led to the analysis and evaluation of the existing scale.

When the results of the markers using the existing scale were analysed statistically, they demonstrated high degrees of reliability/scoring validity. However, the qualitative evidence, extracted from markers of the scripts used in this study, as well as comments from tutor and markers at Unisa, revealed weaknesses regarding the criteria and levels, which did not cover the requirements of the construct. These issues are discussed in the findings of sub- questions 2 and 4.

8.2.2 Sub- question 2

What are the observations of the tutors and markers who use the scale in order to assess examinations and assessments for this module?

Tutors included e-tutors as well as those who tutored at the Parow Regional Centre. Markers were those who marked the 60 scripts during the research process, as well as

the Unisa markers of ENG1501 employed to mark formative and summative assessments.

The following recurring themes were found in the comments of markers at various stages of the process and also arose from the more structured questionnaire sent to tutors and markers.

8.2.2.1 *Number of levels*

Tutors and markers were concerned about the wide range of marks represented by single levels, especially Level 3, which included the pass mark (50%) in the centre of the range without any indication of the cut-off point between a “pass” and a “fail”. Level 2 (56% to 75%) also presented a problem, particularly as regards formative feedback. It was believed that the wide ranges indicated by these levels did not provide sufficient information for the student, and also did not provide enough guidance to the marker. A suggestion was also made to add a further level at the top of the scale for exceptional answers.

8.2.2.2 *Weighting of marks*

Another concern was the weighting of the marks allocated to content/organisation and language use respectively. It was felt that, since this module tests the knowledge and interpretation of a literary text, content should be prioritised above language, especially given the demographic profile of the target student population, the majority of whom were EAL speakers. In addition, the relationship between content and language use was found to be much more intricate and integrated than demonstrated by the marking grid.

8.2.2.3 *Plagiarism*

Extensive plagiarism also presented a problem. The allocation (or lack thereof) of marks for scripts demonstrating plagiarism, and particularly partial plagiarism, also resulted in inconsistent marking, as the current scale provides no penalty for this widespread practice, exacerbated in the ODL context (Minnaar, 2012). It was believed that this omission had a negative impact on the construct validity of the scale as well as on the

concomitant scoring validity (reliability). This is because students can obtain marks for another writer's work, and markers can also mark inconsistently, with some being stricter or more lenient than others in penalising plagiarism.

8.2.2.4 Subjectivity

While it was acknowledged that a measure of subjectivity was unavoidable in developing and applying criteria, tutors and markers considered certain terms such as “shaky” to be too subjective and, thus, open to misinterpretation.

8.2.3 Sub- question 3

What effect, if any, does the distance learning, multilingual and multicultural context have on the perceived and actual validity of the scale?

The following findings were made concerning these issues. Although it was presented as one question to emphasise the close relationship between the two aspects in the case of the target group, each is evaluated separately in the discussion of the findings.

8.2.3.1 The ODL environment

The complex ODL environment in which this research was undertaken was examined at some length in Chapter 2. Stakeholders suggested that improvements should be made to the scale's levels and descriptors to facilitate marking and, very importantly, to assist formative assessment because written feedback (incorporating the mark and remarks) is usually the only contact between the student and the marker. The scale should therefore be clear and accessible to the student, and should be dealt with in the study material, as well as possibly in an introductory lesson by the online tutor. Furthermore, the stressful marking environment, tight deadlines and relative lack of communication between markers in an ODL environment necessitates an assessment scale that is valid and which contains clearly expressed directives and at the same time, is practicable (“user-friendly”).

8.2.3.2 The multicultural and multilingual context

This is an extremely complex topic, especially in the present context where the majority of students are not L1 speakers of the target language, and who speak SAE as an indigenised variety of English, with varying degrees of proficiency.

In a study of the complex South African socio-linguistic situation, Görlach (1998: 108) points out that: “While all informed commentators appear to agree that English will be the dominant language in the years to come, its future norms are much debated”. The challenge is to achieve a balance between the conservative viewpoint that seeks to avoid what it labels “a development towards internal disintegration, frequent miscommunication and, above all, international stigmatization”, and the opposing belief that a “more lenient and realistic attitude would rely on the social prestige of ‘correct’ English but allow for much more variation within the (‘modified’) standard” (Görlach, 1998: 108). In effect, this signals a conflict between descriptive and prescriptive grammatical perspectives. Görlach (1998: 108) favours the adoption of “realistic aims” that stress the instrumental functions of English in the teaching context, and with the aim of teaching English “that is within reach of the learners”. He argues (1998: 108) argues that this practice might lead to “an internal norm (with *lingua franca* uses and intelligibility stressed) and a formal variety closer to international standard”.

Twenty-one years later, there is a greater recognition of indigenised varieties of English (Schneider 2011), but the debate with regard to the degree to which features of SAE should be allowed in formal academic contexts needs to be handled at course development level.

The question thus posed was whether and how a rating scale should reflect these issues in its criteria and/or directives. Participants in the current study expressed some reservations about the extent that a marking scale could address the multicultural and multilingual context. In all three of the grids tested in the current research, the emphasis was on intelligibility as the central criterion of the language component, and it could be argued that this includes the acceptance of entrenched features of SAE, provided that meaning remains clear. Whether the grid should make explicit reference to which features are and are not acceptable is debatable, however, given the diversity and

complexity of SAE. This might lead to the grid becoming over-prescriptive and also unwieldy.

8.2.4 Sub- question 4

What recommendations, principles and insights from other stakeholders can be employed to create an improved scale?

The responses to this secondary question were summarised according to the recurring themes identified in the feedback from stakeholders such as the panel members and other consultants who were contacted in the course of the research. The themes and concerns were as follows.

8.2.4.1 *Number of levels*

Once again, the scale levels were cited as a central concern. Stakeholders were unanimous that the number of levels on the existing scale was inadequate and the range of marks represented by certain levels was too wide. This was a particular problem in the case of Level 3. In agreement with some markers, panel members recommended a further level at the top of the scale in order to reward excellence and encourage students to aim higher than the 75% required for a distinction. This would discourage the perception, as described by a marker, that “75% has become the new 100%”.

8.2.4.2 *Type of scale*

The type of scale was also debated at length, with panel members favouring the two-dimensional scale, although the revised version of the existing scale (Model 1) was also considered to be an improvement. Panel members considered the Likert Scale unsuitable for the given context, as it is too unwieldy and time-consuming to meet the combination of accuracy and practicability demanded by the ODL context.

8.2.4.3 *Weighting of content versus language use*

Stakeholders were concerned about the weighting of content versus language use, especially in the context of assessing assignments on literary texts. They recommended

an assessment approach that prioritised the understanding and interpretation of the literary text above language use. In other words, it should not be possible for a student to obtain a pass mark as a result of competent language use if that student had not studied or understood the text. For instance, one of the markers observed that the most important criterion in a rubric should be whether the candidates address the question and “get it across”, despite language errors.

Panel members were also in agreement with the viewpoint expressed by markers querying the practice of giving two separate marks for an assignment instead of looking at how all the criteria work together. As one of the panel members stated in an email (Appendix G) “I would just hate to give two separate marks and then add them up; to me holism is the essence of the process”.

8.2.4.4 *Formative assessment*

The observation that improvements should be made to the scale levels and descriptors to facilitate marking and, very importantly, to assist formative assessment was again cited, since written feedback (incorporating the mark and remarks) is usually the only contact between the student and the marker. Therefore, the scale should be clear and accessible to the student. Furthermore, the stressful marking environment, tight deadlines and relative lack of communication between markers in an ODL environment necessitated an assessment scale that is valid, contains clearly expressed directives and, at the same time, is practicable, enabling quick but efficient marking.

8.2.5 Sub- question 5

How can the modified or new rating scale be designed and tested to ensure optimum validity?

Two scales were designed by a panel of experts, including the researcher, and tested following the same qualitative and quantitative processes that were employed in the testing of the existing scale. Scripts were marked using both scales and then statistically calibrated. Comments were forthcoming from the markers immediately after the marking, and panel members as well as markers were requested to complete a

questionnaire on both scales at the end of the process. This ensured a multi-pronged and mixed-media approach.

Statistically, both scales obtained a high Cronbach Alpha reliability rating (Table 7.13). Model 1 (the revised scale) scored slightly higher in terms of Item reliability (for the reasons given in Section 7.6.2), but qualitatively, markers showed a strong preference for Model 2 (the two-dimensional scale) and gave convincing reasons for this choice (Section 7.7).

8.3 RECOMMENDATIONS

Since the findings demonstrated a great deal of repetition and overlapping, the recommendations based on the findings of this study have been presented in the following overview.

8.3.1 The validity of the existing scale

As a result of the research process, it was concluded that, while the reliability of the existing scale was high, there were questions about its construct and content validity i.e. the scale did not measure the underlying abilities of interpreting literary texts (reading) and basic academic writing, as stated in the outcomes (Appendix C). The scale also did not cover the specific content of the module, since it was a generic grid shared with other modules which had different outcomes. A scale that is tailor-made to suit the exigencies of a literature module would be more appropriate as it would reflect the skills required by the module.

In response to the findings regarding the validity of the existing scale, the panel recommended that the criteria be amended to suit the specific outcomes of the module, instead of the generic approach, shared unaltered by the rating scales of other modules. Levels should be adapted and increased to allow for more accurate marking. At least one more level should be added to Level 3 of the existing scale since, at present, there is no cut-off point at the 50% mark (indicating the difference between passing and failing).

The scale should also provide directives on penalties to be applied in cases of plagiarism.

8.3.2 The complex ODL, multilingual and multicultural environment

There were some reservations regarding the extent to which the scale could address the challenges of the multicultural and multilingual environment, especially in the complex ODL context. This led to a discussion of the possibility of introducing an alternative module to cater for L2 students, introducing them to the foundations of literary study and to academic essay writing. This is presented merely as a suggestion as it is not directly related to the design of a rating scale for the module ENG1501. However, it was noted that the focus of the existing FAL Grade 12 assessment of literature is on contextual questions and that, therefore, a scaffolding technique in designing assignments should be used to bridge the gap between school and undergraduate literature study, which is based mainly on essay-type questions. An accessible rating scale could support such a scaffolded essay approach.

The panel was of the belief that challenges presented by the diverse multicultural and multilingual target group were exacerbated by the distance learning context, which is characterised chiefly by geographical distance with its concomitant communication and logistical problems. Furthermore, the diversity of the student body raised issues of the role played by the marking grid and the extent to which characteristics of SAE should be incorporated in the marking process. The debate was whether and how to incorporate features of SAE into South African teaching and learning in formative assessment at ODL university institutions. It was argued that the stress placed on intelligibility as a central criterion in the language section of all three grids (i.e. the existing grid, Model 1 and Model 2) should prevent undue penalties for features of SAE varieties. The panel decided against adding explicit directives on the extent to which features of SAE are to be accepted. It was argued that such directives would cause the grid to become unwieldy and too prescriptive (Section 7.7.1.3.2). This issue should be presented in the learning materials and discussed with the marking team in order to ensure inter-rater consistency. The rating scale could also be addressed by means of a podcast or PowerPoint presentation in the online tutor sessions for students, thus incorporating the

ODeL aspect of distance learning, and furthering its aim of ameliorating the problems of distance and communication experienced in this environment.

As described throughout the thesis, and particularly in Chapters 2 and 4, the research context is a unique, extremely complex and very challenging environment. The central challenge is to design a rating scale that is accessible and easy to use but which avoids compromising consequential validity by being the “‘quick fix’ approach” criticised by Steeples *et al.* (2002: 323). This theme overlaps with the discussion of formative assessment in Sections 7.2 and 8.3.6.

8.3.3 Number of levels

As noted in discussing secondary question 1, the wide ranges reflected by the existing scale, particularly Level 3, had a negative impact on the construct validity of the scale. It was thus recommended that more levels be added, while ensuring that the scale did not become ungainly and difficult to follow, especially in the stressful marking situation engendered by the ODL environment. Consequently, Level 3 was split into two at the pass mark. The other levels were adjusted to provide more directives to the markers and formative guidance to the students. Thus, ultimately the panel decided to implement six levels in line with the recommendations of the stakeholders.

8.3.4 Weighting of marks

It was agreed that the weighting of content and language use should be adjusted. Despite the fact that the relative weighting of these features was not indicated in the outcomes (Appendix C), it should not be possible to pass an assignment or examination in a literature module as a result of good language usage but little knowledge and understanding of the prescribed text.

The two-dimensional design (Model 2) was found to offer the best solution to the weighting issue and was recommended by the panel. However, Model 1 was considered to be an improvement on the existing scale. The disadvantage of Model 1 was a bias towards language use at the expense of content knowledge and interpretation. This had a

negative impact on validity. It was thus decided to recommend the two-dimensional scale (Model 2), which adopted an integrated approach to content and language use and prevented bias both in terms of content and language use. This would make it impossible to pass as a result of either content or language use alone.

8.3.5 Plagiarism

The allocation of marks should be adjusted in the case of plagiarism and partial plagiarism. In the final version of Model 2, plagiarised scripts are relegated to Level 6 (“Fail. Seriously at risk”) and those that are partially plagiarised (e.g. a paragraph or numerous sentences) are placed in Level 5 (“Fail. At risk”). It was essential that markers were consistent when penalising plagiarism, and thus it was suggested that the topic be addressed during the markers’ training session.

8.3.6 Subjectivity

It is not possible to eliminate all subjectivity in terms of the rating scale. Even commonly accepted criteria such as “exceptional”, “good”, and “average” lend themselves to a certain degree of subjective interpretation. It can be argued that this is true of all subjects, and is not confined to the Humanities. However, in the case of the existing rating scale, certain terms were considered vague and open to misinterpretation. It was thus recommended that these terms be either omitted or replaced. In the final version of Model 2, the term “shaky” and “very shaky” were replaced by “Fail. At risk”, and “Fail. Seriously at risk”. After some discussion, the term “average” was replaced by “adequate” for the reasons given in Sections 7.4.4 and 7.7.1.2. The argument was that, strictly, technically speaking, “average” reflects a mean, not an “adequate” score. Although other stakeholders argued that this change was unnecessary as the marks and criteria clarify the category, ultimately, the amendment was made in the interests of increased clarity for markers and students.

8.3.7 Formative assessment

The panel adopted the suggestion (as discussed in Section 7.2) that the assessment scale could be included with the scripts, and that markers use this to provide feedback by

underlining or ticking relevant features in need of attention and/or those demonstrating good work. The scale could then be returned to students along with the marked script. It would also be possible to include an electronic copy of the marking grid as part of the feedback when returning assignments submitted electronically. This would provide an effective and relatively easy way of commenting on the student's work. As suggested in Appendix G, another idea would be to familiarise the students with the grid and then give the percentage and category e.g. D4 along with a brief comment. This might be more practicable.

Panel members also considered the use of graphics to make the grid more accessible to students, although no decision was reached because of the difficulty in choosing culturally neutral graphics that would be understood by all groups of the student body. Social media and the use of mobile telephones were considered as valuable means of feedback (Sauder *et al.* 2016: 8). A recommendation was that, given the almost universal use of mobile telephones, the marking grid should be available in an on-screen version, and students could 'google' words or terms unfamiliar to them (Section 5.6).

8.3.8 Type of scale

The two-dimensional design of Model 2 was found to offer the better solution to the issue of weighting and was recommended by the panel. However, Model 1 was considered to be an improvement on the existing scale. The disadvantage of Model 1 was a bias towards language use at the expense of content knowledge and interpretation. This had a negative impact on validity. It was thus decided to recommend Model 2, which adopted an integrated approach to content and language use and prevented bias both in terms of content and language use.

8.4 REPONSES TO THE PRIMARY RESEARCH QUESTIONS

Thus, based on the findings and recommendations of this study, the following responses to the two-fold primary research questions are provided.

- 1. Is the existing assessment scale used for the *Foundations in English Literary Studies* (ENG1501) at Unisa valid in terms of the various aspects of validation and purposes (namely formative assessment, summative assessment, feedback)**

Although the existing assessment scale was valid quantitatively, its qualitative validity was compromised by its generic nature, which prevented it from specifying directives to suit the exigencies of the construct. The wide range represented by the scale levels, and the lack of directives regarding plagiarism also affected validity.

- 2. Depending on the results of the study, how can the existing scale be modified or replaced in order to produce an empirically validated scale for assessing the assignments of the target group?**

This was achieved by using the combined quantitative and qualitative processes described in Chapter 7 and summarised in Section 8.2 and 8.3 of this chapter. The process culminated in the choice of a two-dimensional grid to replace the existing rating scale.

8.5 LIMITATIONS OF THE RESEARCH

One of the limitations of this study was the relatively small size of the sample (60 scripts). However, the research sample was deemed to be sufficient, as it represented the levels indicated on the rating scale, and because the scripts were very carefully and thoroughly scrutinised by several experts, as well as being subjected to the statistical process described in Chapters 5, 6 and 7. Furthermore, this sample size was in line with similar research, which used smaller samples to investigate validity (e.g. Hattingh, 2009).

Another limitation was that the focus of the research on a poetry assignment which consisted of a number of questions (although the assignment was assessed holistically) and that it did not examine later essay-type assignments which covered other literary genres. Although a very small sample of essay-type answers ($n = 4$) was tested in order to address the perceived gap partially, this could be expanded significantly for future research. However, it was believed that, because the poetry assignment tests the given criteria as stated in the outcomes (Appendix C), it should be unnecessary to design a scale for each genre. In fact, it could be argued that a different scale for different genres could confuse markers and students and potentially compromise scoring validity, thus having a negative effect, especially on formative assessment which relies on clear feedback. In addition, Assignment 1 could be seen as a scaffold to the later essay-type assignments and, thus, using the same marking grid makes sense for this reason.

Furthermore, although the research concentrated on feedback from markers and tutors, it did not extend to obtaining input from students, apart from the use of their scripts. It is acknowledged that student feedback, if planned and organised over a considerable period, would have been a rich source of evidence. However, the researcher and panel bore in mind the caution of Simpson and McKay (2013: 25), who question “the idea that rubrics necessarily capacitate students and increase their confidence in approaching assessment tasks”. Findings of a weak correlation between self- and other-assessment were corroborated by Trofimovich *et al.* (2014: 5). The researcher and panel believed that, in order to be effective, a long-term project should be undertaken with a specific focus on students’ feedback on, and interaction with, the scale. Simpson and McKay (2013: 25) note that, in the course of their research, the alignment of marks improved with practice and that, in comparison with first-year students, the revised marks of those in their fourth year of study “were more closely aligned with the score awarded by the lecturer”. This implies “a developing understanding of the expectations contained in the assessment rubric”. It is hoped that this medium can be explored by further research but, for the reasons given, it was felt that the scope of the current study should be limited to input from markers, tutors and lecturers.

For the purposes of this study, the stakeholders of the project were experienced educators and markers who had had extensive interaction with students in a variety of

contexts, including ODL. Thus, they could make authoritative observations about the difficulties experienced by students as they attempted to interpret feedback on their written assignments.

8.6 RECOMMENDATIONS FOR FURTHER STUDY

There are a number of research avenues that can be followed as further study arising from this project.

8.6.1 Research on student input

The current research could be expanded and complemented by a study involving ODL students enrolled for ENG1501 and other literature-based modules. This could be seen as a continuation of the current research, and could take the form of an investigative, qualitative study, including structured and semi-structured interviews, surveys and questionnaires, focused specifically on student input. It could also take the form of case studies, during which individual students' progress could be tracked in the course of a module in order to ascertain the formative value of the assessment scale with particular reference to the techniques used in communicating the assessment criteria. This would be easier in the year-long course envisaged for 2020. The use of technology in formative assessment is a particularly relevant area of research, and yields valuable areas for further studies.

8.6.2 Use of different genres and target groups

A validation study similar to the present research could be conducted to test the two-dimensional scale using essay-type assignments and dealing with literary genres other than poetry. The scale could also be tested on other literary modules, such as the second- and third-year literature courses.

8.7 CONCLUSION

An overview of the findings of the present research were provided in this chapter and recommendations were made for the implementation of the findings by responding to

the research questions posed at the outset of the study. The limitations of the study were then identified, followed by suggestions for further research.

The aim of this thesis was to address the gap caused by the perceived lack of research on assessment process in the case of modules such as that of the target group (ENG1501: *Foundations in English Literary Studies*). The research addressed the challenging environments of ODL, the very diverse student population, and the exigencies of a first-year literature course in this context. It is firmly believed that an accurate, appropriate rating scale aligned to the given construct will contribute to the effectiveness of the assessment process in this module, and act as a scaffold in formative assessment. The process of developing such an assessment scale led to the design of Models 1 and 2, and it is believed that the results addressed the issues raised in the evaluation of the existing scale.

While every effort has been made to verify the validity of the recommended scale, it is acknowledged that no scale is perfect or watertight, and that assessment design and implementation is an ongoing activity that must constantly endeavour to address many complexities. However, it is hoped that the validated rating scale that is the outcome of this study will be used to measure what it is intended to measure, and to fulfil the purpose for which it was intended, namely to provide the necessary features and criteria for summative and formative assessment in the challenging multilingual and multicultural ODL environment.

o-O-o

List of References

- Alderson, J. C. 1991. Bands and scores J. C. In Alderson & North, B. (eds). *Language testing in the 1990s*. London: Macmillan.
- Alderson, J. C. & Banerjee, J. 2002. Language testing and assessment (Part 2). *Language Teaching*, (35):79–113.
- Alderson, J. C., Clapham, C. & Wall, D. 1995. *Language test construction and evaluation*. Cambridge: Cambridge University Press.
- .
- Anastasi, A. 1976. *Psychological testing*. Fourth edition. New York, NY: Macmillan.
- APA (American Psychological Association). 1954. *Standard for educational and psychological testing*. Washington, DC: AERA.
- Asfour, M. 1983. Cultural barriers in teaching literature to Arab students. In E.A Dahiyat & M. H. Ibrahim (eds). Papers from the first conference on the problems of teaching English language and literature at Arab universities. Amman: University of Jordan, 78–92.
- Astika, G. G. 1993. Analytical Assessments of Foreign Students' Writing. *RELC Journal*, 24 (1): 61-72
- .
- Azizi, M. & Majdeddin, K. 2014. On the validity of IELTS writing component: Do raters assess what they are supposed to? *Modern Journal of Language Teaching Methods*, 4(1):337–352.
- Babbie, E. & Mouton, J. 2004. *The practice of social research*. Cape Town: Oxford University Press.

- Bachman, L. F. 1990. *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, L. F. 2004. *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.
- Bachman, L. F. 2005. Building and supporting a case for test use. *Language Assessment Quarterly*, 2(1):1–34.
- Bachman, L. F. & Palmer, A. S. 1996. *Language testing in practice*. Oxford: Oxford University Press.
- Bachman, L. F. & Palmer, A. S. 2010. *Language assessment in practice*. Oxford: Oxford University Press.
- Badal, B. 2016. Pragmatic interpretation: There is a difference in the way that L1 and L2 learners experience the interpretation of a literary text. *Journal for Language Teaching*, 50(2):123–141.
- Bain, J. 2010. Integrating student voice: Assessment for empowerment. *Practitioner Research in Higher Education*, 4(1):14–29.
- Biggs, J. 1999. *Teaching for quality learning at university*. Buckingham: SRHE and Open University Press.
- Blue, G. (ed). 2003. *Developing academic literacy*. Bern: Peter Lang.
- Brindley, G., 1989. The role of needs analysis in adult ESL programme design. In K.R. Johnson (Ed) *The second language curriculum*. New York. Cambridge University Press, 63-78.
- Borsboom, D., Van Heerden, J. & Mellenbergh, G. J. 2003. *Validity and truth*. Retrieved from <http://users.fmg.uva.nl/dborsboorn/BorsboomTroth2003.pdf> [Accessed 10 March 2016].

- Borsboom, D., Mellenbergh, G. H. & Van Heerden, J. 2004. The concept of validity. *Psychological Review*, 111(4):1061–1071.
- Boughey, C. 2007. Educational development in South Africa: From social reproduction to capitalist expansion? *Higher Education Policy*, 20:5–18.
- Boughey, C. 2013. What are we thinking of? A critical overview of approaches to developing academic literacy in South African higher education. *Journal for Language Teaching*, 47(2):25–42.
- Brown, G. T. L., Glaswell, K. & Harland, D. 2004. Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing*, 9(2):105–121.
- Brown, J. D. & Hudson, T. 2002. *Criterion-referenced language testing*. Cambridge: Cambridge University Press.
- Brualdi, A. 1999. *Traditional and modern concepts of validity*. ERIC Digests. Retrieved from <http://www.ericdigests.org/2QQQ-3/validityv.htm> [Accessed 21 April 2014].
- Butler, I. 2006. Integrating language and literature in English studies: A case study of the English 100 course at the University of the North West. Unpublished doctoral thesis. Potchefstroom: North-West University.
- Carter, R. & Long, M. N. 1991. *Teaching literature*. Harlow: Longman.
- Chapelle, C. A. 1999. Validity in language assessment. *Annual Review of Applied Linguistics*, 19:254-272.
- Chapelle, C. A. 2012. Validity argument for language assessment: The framework is simple. *Language Testing*, 29(1):19–27.

- Cheng, G. & Chau, J. 2016. Exploring the relationships between learning styles, online participation, learning achievement and course satisfaction: An empirical study of a blended learning course. *British Journal of Educational Technology*, 47(2):257–278.
- Cheng, L. 2008. Washback, impact and consequences. In E. Shohamy & N.H. Hornberger (eds.). *Encyclopaedia of language and education, Vol. 7: Language testing and assessment*. 2. New York, NY: Springer Science C Business Media, 349–364.
- Cheng, L. & De Luca, C. 2011. Voices from test-takers: Further evidence for language assessment validation and use. *Educational Assessment*, 16:104–122.
- Chokwe, J.M. 2011. Academic writing in English second language contexts: Perceptions and experiences of first year university students and tutors. Unpublished master's dissertation. Pretoria: University of South Africa.
- Cicchetti, D. V. 1994. *Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology*. *Psychological Assessment*, 6(4):284–290.
- Coetzee, M. 2002. *Getting and keeping your accreditation: The quality assurance and assessment guide for education, training and development providers*. Pretoria: Van Schaik.
- Cormier, D. & Siemens, G. 2010. Through the open door: Open courses as research, learning, and engagement. *Educause Review, July/August*: 31–39.
- Council of Europe. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.
- Creswell, J. W. & Clark, V. L. 2007. *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.

- Crocker, L. 2003. Teaching for the test: Validity, fairness, and moral action. *Educational Measurement: Issues and Practice*, 22(3):5–11.
- Cronbach, L. J. 1988. Five perspectives on validity argument. In H. Wainer & H. Braun (eds). *Test Validity*. Hillside, NJ: Lawrence Erlbaum Associates, Inc..
- Cronbach, L. J. & Meehl, P. E. 1955. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302.
- Cumming A., Kantor, R. & Powers, D. 2002. Decision making while assessing ESL/EFL writing: A descriptive framework. *Modern Language Journal*, 86(1):67–96.
- Cureton, E. E. 1951. Validity. In E. F. Lindquist (ed). *Educational measurement*: 621–694. Washington, DC: American Council on Education.
- Department of English Studies. 2013a. *Foundations in English Literary Studies ENG1501: Only study guide for ENG1501*. Pretoria: Unisa Press.
- Department of English Studies. 2013b. *Foundations in English Literary Studies ENG1501: Tutorial letter 101/2014*. Pretoria: Unisa Press.
- Douglas, D. 2000. *Assessing language for specific purposes*. Cambridge: Cambridge University Press.
- Dovey, T. 1994. Making changes without changing: First year courses at five South African universities. *Perspectives in Education*, 15(2):285–298.
- Du Plessis, C. L. 2012. The design, refinement and reception of a test of academic literacy for postgraduate students. Unpublished master's dissertation. University of the Free State.
- Du Plessis, C. 2014. Issues of validity and generalisability in the Grade 12 English Home Language examination. *Per Linguam*, 30(2):1–19.

- Du Plessis, C. & Weideman, A. 2014. Writing as construct in the Grade 12 Home Language curriculum and examination. *Journal for Language Teaching*, 48(2):127–147.
- Durant, A. 1993. Interactive approaches to teaching literature in Hong Kong. In C.J. Brumfit & M. Benton (eds). *Teaching literature: A world perspective*, 150–171. London: Macmillan in association with Modern English Publications and The British Council.
- Ebel, R. L. & Frisbie, D. A. 1991. *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice-Hall.
- Elbow, P. 1996. Writing Assessment in the 21st century: A utopian view. In L. Z. Bloom, D. A. Daiker & E. M. White (eds). *Composition in the twenty-first century: Crisis and change*. Carbondale, IL: Southern Illinois University Press.
- Elder, C. 2005. Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly*, 2(3):175–196.
- Ellis, R. 1994. *The study of second language acquisition*. Oxford: Oxford University Press.
- Elton, L. 1993. University teaching: A professional model for quality. In R. Ellis (ed). *Quality Assurance for university teaching*. Buckingham: The Society for Research into Higher Education and Open University Press, 133–146.
- Falvey, P. & Kennedy, P. (eds). 1997. *Learning language through literature: A source book for teachers of English in Hong Kong*. Hong Kong: Hong Kong University Press.
- Frederiksen, J. R. & Collins, A. 1989. A systems approach to educational testing. *Educational Researcher*, 18(9):27–32.

- Fulcher, G. 1996. Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing*, 13(2):208-238.
- Fulcher, G. 1999. The communicative legacy in language testing. *Applied Linguistics*, 28(4):483-497.
- Fulcher, G. & Davidson, F. 2007. *Language testing and assessment: An advanced resource book*. Abingdon: Routledge.
- Fung, D. 2017. *A connected curriculum for higher education*. London: UCL Press.
- Garson, G. D. 2006. *Validity. Statnotes: Topics in multivariate analysis*. Retrieved from <http://www2.chass.ncsu.edu/garson/pa765/validity.htm> [Accessed 10 April 2017].
- Geisler, C. 1994. *Academic literacy and the nature of expertise: Reading, writing and knowing in academic philosophy*. Hillsdale, NJ: Lawrence Erlbaum.
- Ghamarian, D., Motallebzadeh, K. & Fatemi, M. A. 2014. Investigating the relationship between the washback effect of IELTS test and Iranian IELTS candidates' life skills. *Journal of Language and Linguistic Studies*, 10(1):137–152.
- Gibbs, G. 2006. Why assessment is changing. In C. Bryan & K. Clegg (eds). *Innovative assessment in higher education*. Abingdon: Routledge: 11–22.
- Golder, K., Reeder, K. & Fleming, S. 2009. Determination of appropriate IELTS band score for admission into a program at a Canadian post-secondary polytechnic institution. In J. Osborne (ed). *International English Language Testing System (IELTS) research reports*. 10:1–25.
- Gorin, J. S. 2007. Reconsidering issues in validity theory. *Educational Researcher*, 36(8):456–462.
- Görlach, M. 1998. *Even more Englishes 1996-1997*. Philadelphia: John Benjamins

- Green, A. 1998. *Verbal protocol analysis in language testing research: A handbook*. Cambridge: Cambridge University Press.
- Greene, J. C. 2007. *Mixed methods in social inquiry*. Volume 9. New York, NY: Wiley.
- Guildford, J. P. 1946. New standards for test evaluation. *Educational and Psychological Measurement*, 6(5):427–439.
- Guion, R. M. 1980. On trinitarian doctrines of validity. *Professional Psychology*, 11(3):385–398.
- Gunawardena, C .N. & McIsaac, M. S. 2004. Distance education. In D.H. Jonassen (ed.). *Handbook of research on educational communications and technology*. Mahwah, NJ: Lawrence Erlbaum, 355–395.
- Gustke, C. 2010. E-learning hits barriers to expansion. *Education Week*, 29(30):S9.
- Hamp-Lyons, L. 1990. Second language writing: Assessment issues. In B. Kroll (ed). *Second language writing: Research insights for the classroom*. Cambridge: Cambridge University Press.
- Hamp-Lyons, L. 1991. Scoring procedures for ESL contexts. In L. Hamp-Lyons (ed.). *Assessing second language writing in academic contexts*. Norwood, NJ: Ablex.
- Hamp-Lyons, L. 1995. Rating normative writers – the trouble with holistic scoring. *TESOL Quarterly*, 29(4):750–762.
- Harker, M. & Koutsantoni, D. 2005. Can it be as effective? Distance versus blended learning in a web-based EAP programme. *ReCALL*, 17(2):197–216.
- Haswell, R. & Wyche-Smith, S. 1994. Adventuring into writing assessment. *College Composition and Communication*, 45(2):220–236.

- Hathaway, J. 2015. Developing that voice: Locating academic writing tuition in the mainstream of higher education. *Teaching in Higher Education*, 20(5):506–517.
- Hattingh, K. 2009. The validation of a rating scale for the assessment of compositions in ESL. Unpublished doctoral thesis. Potchefstroom: North-West University.
- Hawkey, R. 2006. *Impact theory and practice: Studies of the IELTS Test and Progetto Lingue 2000*. Cambridge: Cambridge University Press.
- Hawkey, R. & Barker, F. 2004. Developing a common scale for the assessment of writing. *Assessing Writing*, 9(2):122–159.
- Henning, G. 1987. *A guide to language testing*. Cambridge, MA: Newbury House.
- Henning, G. 1991. Issues in evaluating and maintaining an ESL writing assessment program. In L. Hamp-Lyons (ed). *Assessing second language writing in academic contexts*. Norwood, NJ: Ablex.
- Hill, J. 1986. *Using literature in language teaching*. London: Macmillan
- Howie, S., Venter, E., Van Staden, S., Zimmerman, L., Long, C., Du Toit, C., Scherman, V. & Archer, E. 2008. *PIRLS 2006 summary report: South African children's reading achievement*. Pretoria: Centre for Evaluation and Assessment, University of Pretoria.
- Howie, S. J., Combrinck, C., Roux, K., Tshele, M., Mokoena, G. M. & McLeod Palane, N. 2017. *PIRLS LITERACY 2016: South African highlights report*. Pretoria: Centre for Evaluation and Assessment, University of Pretoria.
- Huang, J. 2009. Factors affecting the assessment of ESL students' writing. *IJAES*, 5(1).
- Hudson, T. 2005. Trends in assessment scales and criterion-referenced language assessment. *Annual Review of Applied Linguistics*. 25:205-227.

- Hughes, A. 1989. *Testing for language teachers*. Cambridge: Cambridge University Press.
- Hughes, A. 2003. *Testing for language teachers*. Cambridge: Cambridge University Press.
- Huot, B. 1990. Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication*, 41(20):201–213.
- Hyland, F. 2001. Providing effective support: Investigating feedback to distance language learners. *Open Learning*, 16(3):233–247.
- IELTS 2007. *IELTS handbook*. University of Cambridge: ESOL examinations.
Retrieved from [http://www.cambridgeesol.org/assets/pdf/resources/IELTS Handbook.pdf](http://www.cambridgeesol.org/assets/pdf/resources/IELTS%20Handbook.pdf) [Accessed 19 April 2014].
- Jacobs, H., Zjnggraf, S. A., Wormuth, D. R., Hartfield, V. F. & Hughey, J. B. 1981. *Testing ESL compositions: A practical approach*. Rowley, MA: Newbury House.
- Johnson, M., Mehta, S. & Rushton, N. 2015. Assessment, aim and actuality: Insights from teachers in England about the validity of a new language assessment model. *Pedagogies: An International Journal*, 10(2):128–148.
- Jones, C., Asensio, M. & Goodyear, P. 2000. Networked learning in higher education: Practitioners' perspectives. *AIT-J*, 8(2):18–28.
- Jones, N. 2001. Reliability as one aspect of test quality. *Research Notes*, 4:2–5.
- Kachru, B. B. 1985. Standards, codification, and sociolinguistic realism: the English language in the outer circle. In R. Quirk & H. Widdowson (eds). *English in the World: Teaching and Learning of Language and Literature*. Cambridge: Cambridge University Press: 11-30.

- Kachru, B. B. 1990. World Englishes and applied linguistics. *World Englishes*, 9(1):3–20.
- Kane, M. 1992. An argument-based approach to validity. *Psychological Bulletin*, 112:527–535.
- Kane, M. 2004. Certification testing as an illustration of argument-based validity. *Measurement*, 2(3):135–170.
- Kane, M. 2006. Validation. In R. Brennan (ed). Educational measurement. Fourth edition. Westport, CT: American Council on Education and Praeger, 17–64.
- Kane, M. 2012. Validating score interpretations and uses: Messick Lecture, Language Testing Research Colloquium, Cambridge, April 2010. *Language Testing*, 29, 3–17.
- Kane, M. 2013. Validity and fairness in the testing of individuals. In M. Chatterji (ed), *Validity and test use*, Bingley: Emerald, 17-53..
- Kane, M. 2017. Loosening psychometric constraints on educational assessments. *Assessment Education: Principles, Policy and Practice*, 24(3):447–453.
- Kane, M. & Bridgeman, B. 2017. Research on validity theory and practice at ETS. *Advancing human assessment, methodology of educational measurement and assessment*: 489–552.
- Kane, M., Crooks, T. J. & Cohen, A. S. 1999. Validating measures of performance. *Educational Measurement: Issues and Practice*, 18(2):5–7.
- Knoch, U. & Elder, C. 2013. A framework for validating post-entry language assessments (PELAs). *Papers in Language Testing and Assessment*, 2(2):48–56.

- Kraemer, A. 2008. Formats of distance learning. In S. Goertler & P. Winke (eds). *Opening doors through distance language education: Principles, perspectives and practice*. San Marcos, TX: CALICO, 11–42.
- Lamb, S. & Simpson, Z. 2011. Students' expectations of feedback given on draft writing. *Per Linguam*, 27(1):44–55.
- Landy, F. J. 1986. Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist*, 41:1183–1192.
- Lane, S. 1999. *Validity evidence for assessments*.
www.nciea.org/publications/ValidityEvidenceLane99.pdf [Accessed 24 March 2016].
- Lephalala, M. & Makoe, M. 2012. The impact of socio-cultural issues for African students in the South African distance education context. *Journal of Distance Education*, 26(1):1–10.
- Lephalala, M. & Pienaar, C. 2008. An evaluation of markers' commentary on ESL students' argumentative essays in an ODL context. *Language Matters*, 39(1):66–87.
- Levy, P. & Petrulis, R. 2012. How do first year university students experience inquiry and research, and what are the implications for the practice of inquiry-based learning? *Studies in Higher Education*, 37(1):85–101.
- Linacre, J. M. 2006. *FACETS Rasch Measurement Computer Program Facets for Windows, version 3.61.0*. Chicago, IL: Winsteps.com.
- Lissitz, R. W. & Samuelson, K. 2007. A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, 36(8):463–469.
- Loevinger, J. 1957. Objective tests as instruments of psychological theory. *Psychological Reports Monograph*, 3(9):635–694.

- Lumley, T. 2002. Assessment criteria in a large-scale writing-test: What do they really mean to the raters? *Language Testing*, 19(3):246–276.
- Lumley, T. & Brown, A. 2005. Research methods in language testing. In E. Hinkel (ed). *Handbook of research in second language teaching and learning*. Mahwah, NJ: Lawrence Erlbaum.
- Luoma, S. 2004. *Assessing speaking*. Cambridge: Cambridge University Press.
- Lydster, C. & Brown, S. 2017. The value of Post-Entry Language Assessment (PELA): Outcomes from a first semester undergraduate subject. *Journal of Academic Language & Learning*, 11(1):A39–A57.
- Macdonald, C. & Pinheiro, M. 2012. Vygotskian methods of teaching and learning in the English classroom: The case of grammar. *Journal for Language Teaching*, 46(1):88–102.
- Machet, M. & Pretorius, E. 2008. The impact of storybook reading on emergent literacy: Evidence from poor rural areas in KwaZulu-Natal, South Africa. *Mousaion*, Special issue: 261–289.
- Mahlangu, V. 2016. Assuring quality in ODL through Ubuntu. *Open distance learning (ODL)*: 107-118.
- Makhanya, M. 2016. Foreword. *Open distance learning (ODL)*:vii –viii.
- Maley, A. 1989. A comeback for literature? *Practical English Teaching*, 10(1):59.
- Marcoulides, G. A. 2004. Conceptual debate in evaluating measurement procedures. *Measurement: Interdisciplinary Research and Perspectives*, 2(3):182–184.
- Marshall, B. 2016. The politics of testing. *English in Education*, 1:1–17.

- McColly, W. 1970. What does educational research say about judging of writing? *Journal of Educational Research*, 64(4): 148 -15 6.
- McKenna, S. 2007. Assessment rubrics as a means of making academic expectations explicit. *The Journal of Independent Teaching and Learning*, 2:22–30.
- McLoughlin, C. 2001. Inclusivity and alignment: Principles of pedagogy, task and assessment design for effective cross-cultural online learning. *Distance Education*, 22(1):7–29.
- McNamara, T. 1996. *Measuring second language performance*. New York, NY: Longman.
- McNamara, T. 2000. *Language testing*. Oxford: Oxford University Press.
- McNamara, T. 2003.. Language testing in practice: designing and developing useful language tests. *Language Testing*, 20(4):466-473.
- McNamara, T. & Roever, C. 2006. *Language testing: The social dimension*. In R. Young (ed.). *Language learning monograph series*. Oxford: Blackwell.
- McRae, J. 1991. Applying the buzz words. *British Book News*, July: 432–437.
- Messick, S. 1975. The standard problem: Meanings and values in measurement and evaluation. *American Psychologist*, 3(10):955–966.
- Messick, S. 1980. Test validity and the ethics of assessment. *American Psychologist*, 35(11):1012–1027.
- Messick, S. 1989. Validity. In R. L. Linn (ed). *Educational measurement*. Third edition: New York, NY: Macmillan, 13-103.
- Messick, S. 1992. Validity of test interpretation and use. In M. C. Alkrm (ed). *Encyclopaedia of educational research*. New York, NY: Macmillan.

- Messick, S. 1994. The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13–23.
- Messick, S. 1996. Validity and washback in language testing. *Language Testing*, 13(4):241–257.
- Minnaar, A. 2012. A framework for controlling dishonesty in open and distance learning (ODL) IN Higher Education. *Journal of Teaching and Education*, 1(3):1–13 (2012)
- Moore, M. & Kearsley, G. 2011. *Distance education: A systems view of online learning*. New York, NY: Wadsworth.
- Moss, P. 2007. Reconstructing validity. *Educational Researcher*, 36(8):470–476.
- Murday, K., Ushida, E. & Chenoweth, A.N. 2008. Learners' and teachers' perspectives on language online. *Computer Assisted Language Learning*, 21(2):125–142.
- Ndaba, S. 2005. *Halos and horns: Reliving constructions of matric performance in the South African education system. Matric: What should be done?* Paper presented at the Umalusi and Centre for Higher Education Transformation Seminar, 23 June. Retrieved from <http://www.umalusi.org.za/ur/research/ndaba.pdf> [Accessed 10 August 2015].
- North, B. & Schneider, G. 1998. Scaling descriptors for language proficiency scales. *Language Testing*, 15(2):217–263.
- Ntuli, C. D. & Pretorius, E. J. 2005. Laying foundations for academic language competence: The effects of storybook reading on Zulu language, literacy and discourse development. *Southern African Linguistics and Applied Language Studies*, 23(1):91–109.

- NWU (North-West University). Retrieved from www.distance.nwu.ac.za [Accessed 14 February 2017].
- O'Sullivan, B. 2000. Toward a model of performance in oral testing. Unpublished doctoral thesis. Place: University of Reading.
- O'Sullivan, B. 2006. *Issues in business English testing: The BEC revision project*. Cambridge: Cambridge University Press.
- Paran, A. 1998. Helping learners to create and own literary meaning in the ELT classroom. *Ideas: The APIGA Magazine*, 1:6–9.
- Patterson, R. & Weideman, A. 2013. The typicality of academic discourse and its relevance for constructs of academic literacy. *Journal for Language Teaching*, 47(1):107–123.
- Pienaar, C. 2005. Shared assessment: Empowering student writers. *Language Matters*, 36(2):193–204.
- Pitsoe, V. & Letseka, M. 2016. Ubuntu driven ODL student assessment. In M. Letseka (ed). *Open distance learning (ODL)*:93–104.
- Popham, W.J. 1981. *Modern educational measurement*. Englewood Cliffs, NJ: Prentice Hall.
- Pretorius, E. J. 2008. What happens to literacy in (print) poor environments? Reading in African languages and school language policies. In *Proceedings of CentRePol/Ifas Workshop*, University of Pretoria, 29 March 2007. Johannesburg: Institut Français d'Afrique du Sud, 60–88.
- Pretorius, E .J. & Currin, S. 2010. Do the rich get richer and the poor poorer? The effects of an intervention programme on reading in the home and school language in a high poverty multilingual context. *International Journal of Educational Development*, 30:67–76.

- Pretorius, E. J. & Ribbens, I. R. 2005. Reading in a disadvantaged high school: Issues of accomplishment, assessment and accountability. *South African Journal of Education*, 25(3):139–147.
- Prodromou, L. 2000. Reason not the need: Shakespeare in ELT. *IATEFL Issues*, 156:2–3.
- Rambiritch, A. 2013. Validating the Test of Academic Literacy for Postgraduate Students (TALPS). *Journal for Language Teaching*, 47(1):175–193.
- Read, J. 2010. Researching language testing and assessment. In B. Paltridge & A. Phakiti (eds). *Continuum companion to research methods in applied linguistics*. London: Continuum International, 286–300.
- Reckase, M. D. 2017. *A tale of two models: Sources of confusion in achievement testing*. Research report. Princeton NJ: Educational Testing Service.
- Relearning by Design. 2000. Retrieved from *Rubricsampler*. <http://www.relearning.org/resources/PDF/rubric> [Accessed 7 April 2014].
- Rennie, F. 2007. Understanding practitioners perspectives of course design for distributed learning. *European Journal of Open, Distance, and E-Learning*. Retrieved from <http://www.eurodl.org/?article=287> [Accessed 3 April 2017].
- Rozeboom, W. W. 1966. *Foundations of the theory of prediction*. Homewood, IL: Dorsey.
- Ruth, D. 2001. Unsettle the mainstream. *Getting ahead: Supplement on higher education, Mail & Guardian*, 2–8 February: 1.
- Ryan, J. M. 2002. Issues, strategies and procedures for applying standards when multiple measures are employed. In G. Tindall & T. M. Haladyna (eds). *Large-*

scale assessment programs for all students: Validity, technical adequacy and implementation. Mahwah, NJ: Lawrence Erlbaum.

Saba, F. 2000. Research in distance education: A status report. *International Review of Research in Open and Distance Learning*, 1(1):1–9.

Saville, N. 2001. Test development and revision. *Research Notes*, 4:5–8.

Scharton, M. 1996. The politics of validity. In E.M. White, W. D. Lutz & S. Kamusikiri (eds). *Assessment of writing: Politics, policies, practices*. New York, NY: Modern Language Association of America.

Schilling, S. G. 2004. Conceptualising the validity argument: An alternative approach. *Measurement: Interdisciplinary Research and Perspectives*, 2(3):178–182.

Schneider, E. W. 2011. *English around the World: An Introduction*. Cambridge: Cambridge University Press.

Shanahan, D. 1997. Articulating the relationship between language, literature, and culture: Toward a new agenda for foreign language teaching and research. *The Modern Language Journal*, 81(2):164–174.

Shandu, T. 2017. Designing and implementing mobile-based interventions for enhancing English vocabulary in ODL. Unpublished doctoral thesis. Pretoria: University of South Africa.

Shaw, S. D. and Jordan, S. 2002. CELS Writing: test development and validation activity. *Research Notes*, 9:10-13.

Shaw, S. D. & Weir, C. 2007. *Examining writing: Research and practice in assessing second language writing*. Studies in Language Testing. Cambridge: Cambridge University Press.

- Shepard, L. 1993. Evaluating test validity. *Review of Research in Education*, 19(1):405–450.
- Shohamy, E. 2001. *The power of tests: A critical perspective on the uses and consequences of language tests*. Harlow: Longman/Pearson.
- Shohamy, E. 2005. Language policy: Hidden agendas and new approaches. *Language Policy: Hidden Agendas and New Approaches*. 1-185.
- Shohamy, E. 2010. Introduction to Volume 7: Language testing and assessment. In E. Shohamy & N.H. Hornberger (eds.). *Encyclopaedia of Language and Education*. New York, NY: Springer.
- Sieborger, R. 2004. *Transforming assessment: A guide for South African teachers*. Cape Town: Juta.
- Silva, P. 2011. The lexis of South African English: Reflections of a multilingual society. In E Schneider (ed). *Englishes around the world*,. 2. Amsterdam/Philadelphia.
- Simpson, Z. & McKay, T. M. 2013. Assessment rubrics: Artefacts that speak in tongues? *Per Linguam*, 29(1):15–32.
- Solé, C. R. & Hopkins, J. 2007. Contrasting two approaches to distance language learning. *Distance Education*, 28(3):351–370.
- Spencer, B. 1997. Responding to student writing: Power and role relations. *Per Linguam*, 13(1):39–51.
- Spencer, B. 1998. Responding to student writing: Strategies for a distance-teaching context. Unpublished doctoral thesis. Pretoria: University of South Africa.
- Spencer, B. 2005. Responding to student writing: A taxonomy of response styles – when language (accuracy) matters too much. *Language Matters*, 36(2):205–223.

- Spencer, B. 2009. Aligned assessment in support of high-level learning: A critical appraisal of an assignment for a distance-teaching context. *Journal for Language Teaching*, 43(2):102–109.
- Spencer, B., Lephalala, M. & Pienaar, C. 2005. Improving academic proficiency in open distance learning through contact interventions. *Language Matters*, 36(20):224–242.
- Steeple, C., Jones, C. & Goodyear, P. 2002 Beyond e-learning: A future for networked learning. In C. Steeples & C. Jones (eds). *Networked learning: Perspectives and issues. Computer supported cooperative work*. London: Springer.
- Taylor, L. 2002. *IELTS writing test revision steering group discussion paper*. Internal report. Cambridge: UCLES.
- Tennant, M., McMullen, C. & Kaczynski, D. 2010. *Teaching, learning and research in higher education - A critical approach*. London: Routledge.
- Teddlie, C. & Tashakkori, A. 2009. *Foundations of mixed methods research: Integrating quantitative and qualitative techniques in the social and behavioral sciences*. Thousand Oaks, CA: Sage.
- Terre Blanche, M. & Durrheim, K. 1999. *Applied methods for the social sciences*. Cape Town: University of Cape Town Press.
- Thang, S. 2005. Comparing approaches to the studying of Malaysian distance learners and on-campus learners: Implications to distance education. *Turkish Online Journal of Distance Education*, 6(2):70–86.
- Thompson, P. International English Language Testing System (IELTS) Research Reports 2009, 9 [online]. [Canberra]: British Council and IELTS Australia, 2009. [Canberra]:

- Trofimovich, P., Isaacs, T., Kennedy, S., Saito, K., & Crowther, D. 2014. Flawed self-assessment: Investigating self- and other-perception of second language speech. *Bilingualism: Language and Cognition*, 19, 1–19.
- Toulmin, S.E. 2003. *The use of argument*. 2nd ed. Cambridge: Cambridge University Press.
- Tsushima, R. 2015. Methodological diversity in language assessment research: The role of mixed methods in classroom-based language assessment studies. *International Journal of Qualitative Methods*, 14(2):104–121.
- Turner, C. E. 2000. Listening to the voices of rating scale developers: Identifying salient features for second language performance assessment. *Canadian Modern Language Review*, 56(4). Retrieved from <http://www.utpiouinals.com/product/cnik/564/564-Tuner.html> [Accessed 2 July 2016].
- Van der Slik, F. & Weideman, A. 2008. Measures of improvement in academic literacy. *Southern African Linguistics and Applied Language Studies*, 26(3):363–378.
- Van der Walt, J. L. 2012. The meaning and uses of language test scores: An argument-based approach to validation. *Journal for Language Teaching*, 46(2):489–546.
- Van der Walt, J. L. & Steyn, H. S. 2007. Pragmatic validation of a test of academic literacy. *Ensovoort*, 11(2):138–153.
- Vorobel, O. & Kim, D. 2012 Language teaching at a distance: An overview of research. *CALICO Journal*, 29(3):548–562.
- Vygotsky, L. S. 1978. *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.
- Vygotsky, L. S. 1986. *Thought and language*. Cambridge, MA: The MIT Press.

- Wang, Y., Peng, H., Haung, R., Hou, Y. & Wang, J. 2008. Characteristics of distance learners research on relationships of learning motivation, learning strategy, self-efficacy, attribution and learning results. *Open Learning*, 23(1):17–28.
- Ward-Cox, M. 2012. A critical review of language errors in the writing of distance education students. Unpublished master's dissertation. Pretoria: University of South Africa.
- Weideman, A. 2003. Towards accountability: A point of orientation for post-modern applied linguistics in the third millennium. *Literator*, 24(1):83–102.
- Weideman, A. 2006. Transparency and accountability in applied linguistics. *Southern African Linguistics and Applied Language Studies*, 24(1):71–86.
- Weideman, A. 2009. Constitutive and regulative conditions for the assessment of academic literacy. *Southern African Linguistics and Applied Language Studies*, 27(3):235–251.
- Weideman, 2012. Validation and validity beyond Messick. *Per Linguam*, 28(2):1–14.
- Weideman, A. & Van Rensburg, C. 2002. Language proficiency: Current strategies, future remedies. *SAALT Journal for Language Teaching*, 36(1/2):152–164.
- Weigle, S. C. 2002. *Assessing writing*. Cambridge: Cambridge University Press.
- Weir, C. J. 1990. *Communicative language testing*. New York, NY: Prentice Hall.
- Weir, C. J. 2005. *Language testing and validation: An evidence-based approach*. Oxford: Palgrave Macmillan.
- Weir, C. J. & Shaw, S. D. 2005. Establishing the validity of Cambridge ESOL Writing tests: towards the implementation of a socio-cognitive model for test validation. *Research Notes*, 21:10-14.

- White, C. (2006). The distance learning of foreign languages. *Language Teaching* 39.4, 247–264.
- White, C. (2008). Language learning strategies in independent language learning: An overview. In T. Lewis & S. Hurd (eds.), *Language learning strategies in independent settings*. Clevedon, UK: *Multilingual Matters*, 3–24.
- White, C. 2008. *Language learning strategies in independent language learning: An overview*. Cambridge: Cambridge University Press.
- White, C. 2014. Thinking allowed. The distance learning of foreign languages: A research agenda. *Language Teaching*, 47(4):538–553.
- Widdowson, H. G. 1973. Literature and its communicative value. *English Academy of Southern Africa*, 17–32.
- Widdowson, H. G. 1974. Literary and scientific uses of English. *ELT Journal*, 28(4):282–292.
- Widdowson, H. G. 1975. *Stylistics and the teaching of literature*. Harlow: Longman.
- Widdowson, P. (ed). 1982. Introduction: The crisis in English studies. *Re-reading English*. London: Methuen.
- Widdowson, H. G. 1983. Talking shop: On literature and ELT. *ELT Journal*, 37(1):30–35.
- Widdowson, H. G. 1984. *The use of literature: Explorations in applied linguistics*. Oxford: Oxford University Press.
- Widdowson, H. G. 1992. *Practical stylistics: An approach to poetry*. Oxford: Oxford University Press.

- Wildsmith-Cromarty, R. & Steinke, K. 2014. The write approach. Can R2L help at tertiary level? *Per Linguam*, 30(1):38–54.
- Wolfe, E., Kao, C. & Ranney, M. 1998. Cognitive differences in proficient and non-proficient essay scorers. *Written Communication*, 15(4):469–482.
- Young, D. J. 2008. An empirical investigation of the effects of blended learning on student outcomes in a redesigned intensive Spanish course. *CALICO Journal*, 26(1):160–181.
- Ysal, H. H. 2010. A critical review of the IELTS writing test. *ELT Journal*, 64(3):314–320.

Appendices

Appendix A: Participation and Informed Consent Leaflet

Researcher's name: Maxine Welland Ward-Cox
Student Number: 2311194
Department of English
University of South Africa

TOPIC: Validation of a rating scale for distance education student essays in a literature-based module.

Dear Participant

I am a doctoral student in the Department of English, University of South Africa. You are invited to participate in our research project regarding validation of a rating scale for student writing in a distance education environment.

This letter contains information to help you with your decision to take part in this study. Please read through the letter carefully in order to make an informed decision. If the information is unclear or if you have any other questions, do not hesitate to ask me. You should not agree to take part in this study unless you fully understand the content of this letter.

NATURE AND PURPOSE OF THIS STUDY

The aim of this study is to determine whether the current rating scale of the module ENG1501 (Foundations in English Literary Studies) is valid within a distance education context.

You, as a participant, are a very important source of information on determining whether the scale is appropriate for both formative and summative assessment in the given context. You are an indispensable and worthy partner in this research. Your rights and interests will be respected at every stage and level of research.

EXPLANATION OF PROCEDURE TO BE FOLLOWED

Students' assignments will be benchmarked and then marked by at least two markers to ascertain the validity of the assessment instrument. These assignments will be randomly selected from those submitted for marking, and confidentiality will be ensured by blocking out the student's name and assigning a number to the script. Students will be informed that their assignments have been chosen for the project and will be asked to sign the consent form. Other activities will involve students, tutors and language experts contributing to discussions, completing questionnaires, participating in interviews and making recommendations to improve the validity of the assessment instrument. If

necessary, another rating scale will be proposed. The new or improved instrument will then be piloted and finally proposed as an alternative to the current scale.

By completing and returning the research questionnaire, you give consent that the information received can be used for the research. During this process, the researcher will be available to answer questions you might have regarding the questionnaire. Follow up interviews may be held to elicit further information from you that may add to the data collected.

RISK AND DISCOMFORT INVOLVED

Except for the time it takes to write the assignment, there is no known discomfort or inconvenience related to this study. Your time and active participation in this study is highly appreciated and invaluable to its successful completion.

POSSIBLE BENEFITS OF THIS STUDY

Writing assignments comprises a significant part of the work while studying at Unisa and students not only receive new knowledge in a subject but also learn a new language – an academic and subject-specific language. The difficulty is that students not only have to come to terms with academic language and writing but also have to do so in a distance education context. The results of this study will be used design a valid, fair writing scale that will measure what it is supposed to measure, and expresses criteria as clearly and as unambiguously as possible, given the challenging context of the DE context.

WHAT ARE YOUR RIGHTS AS A PARTICIPANT IN THIS STUDY?

Your participation in this study is entirely voluntary and you can refuse to participate or stop at any time.

COMPENSATION

Your participation is voluntary. No compensation or contribution towards your transport expenses or other expenses will be given for your participation.

CONFIDENTIALITY

All information obtained during the course of this study is strictly confidential. Data that may be reported in scientific journals will not include any information that can identify you as a participant in this study.

HAS THE STUDY RECEIVED ETHICAL APPROVAL?

The study proposal was submitted to The Higher Degrees Committee of the Department of English Studies in the College of Human, the University of South Africa. Written approval to conduct the study has been approved.

INFORMATION AND CONTACT PERSON

If you have any questions during this study, please do not hesitate to approach the researcher.

Researcher: Maxine Ward-Cox

Contact details: maxibob@telkomsa.net

Supervisor: Prof Brenda Spencer

Appendix A1: Research Permission

RESEARCH PERMISSION SUB-COMMITTEE OF SRIHDC 16 March 2015

Dear Mrs Maxine Ward-Cox, Decision: Research Permission Approval

Name: Mrs Maxine Ward-Cox College of Human Sciences Department of English
UNISA maxibob@telkomsa.net (021) 557-3314/083 285 5971
Supervisor: Prof Brenda Spencer spencb@unisa.ac.za

A study titled: "Validation of a rating scale for distance education university student essays in a literature-based module."

Your application regarding permission to conduct research involving UNISA data in respect of the above study has been received and was considered by the Research Permission Subcommittee (RPSC) of the UNISA Senate Research and Innovation and Higher Degrees Committee (SRIHDC) on 12 March 2015.

It is my pleasure to inform you that permission has been granted for your study, for the period between 16 March 2015 and 31 December 2018, to access and use the essays scripts of students registered for ENG1501 through the assistance of the Module coordinator.

Note: The reference number 2015_RPSC_016 should be clearly indicated on all forms of Ref #: 2015_RPSC_016 Mrs Maxine Ward-Cox Student #: 02311194 Staff #:

communication with the intended research participants.

We would like to wish you well in your research undertaking.

Kind regards,

PROF L LABUSCHAGNE EXECUTIVE DIRECTOR:
RESEARCH

Appendix A2: Participant's Consent Form

TOPIC:

I.....hereby agree to participate in a study with the title “ Validation of a rating scale for distance education student essays in a literature-based module”. I hereby acknowledge that I am participating in this research voluntarily, and am aware that I may withdraw from the research at any time. I agree that the results be recorded on condition that anonymity and confidentiality will be maintained.

I understand that agreeing to take part means that I am willing to:

- Allow my written assignment to be assessed, provided that confidentiality is maintained
- Be informed about the research results.

The purpose of this research is to fulfill the requirement for the PhD Degree in Teaching English to Speakers of Other Languages.

I understand that the information provided by me shall remain confidential:

- My participation is voluntary,
- I can choose not to participate in part or all of the study, and
- I can withdraw at any stage without being penalized or disadvantaged in any way.

Name of participant

Signature

Date

Name of researcher

Student number

Signature.....

Date

Appendix A3: Deelnemer se Toestemmingsvorm

ONDERWERP:

Ek,, stem hiermee in om deel te neem aan die studie met die titel “Validation of a rating scale for distance education university student essays in a literature-based module”. Ek erken hiermee dat ek vrywillig aan hierdie navorsing deelneem, en dat ek bewus is daarvan dat ek ter enige tyd my aan die studie kan onttrek. Ek gee my toestemming dat die resultate opgeteken kan word op voorwaarde dat dit anoniem bly en dat vertroulikheid behoue bly.

Ek verstaan dat my toestemming om deel te neem beteken dat ek bereid is:

- Om toe te laat dat my geskrewe werkstuk geassesseer word, op voorwaarde dat vertroulikheid behoue bly
- Om ingelig te word aangaande die resultate van die navorsing.

Die doel van hierdie studie is om te voldoen aan die vereistes van die PhD-graad in ‘Teaching English to Speakers of Other Languages’.

Ek verstaan dat die inligting wat ek sal verskaf vertroulik sal bly:

- My deelname is vrywillig,
- Ek kan verkies om nie deel te neem aan ’n deel van die studie of die hele studie nie
- Ek kan op enige stadium aan die studie onttrek sonder dat ek op enige manier geenaliseer sal word of te na gekom sal word.

Naam van deelnemer

Handtekening

Datum

Naam van navorser

Studentenommer

Handtekening.....

Datum

Appendix A4: Ifomu Yemvume Yomthabathi Nxaxheba

ISIHLOKO:

Mna.....ndiyavuma ukuthabatha inxaxheba kuphononongo elinetatile " Validation of a rating scale for distance education university student essays in a literature-based module". Ndilapha ndiyavuma ukuba ndithabatha inxaxheba kolu phononongo ngokuzithandela, kwaye ndiyazi ukuba ndingarhixa kuphando nangaliphi ixesha. Ndiyavuma ukuba iziphumo zishicilelwe ngaphandle kwegama kwaye ubumfihlo buza kugcinwa.

Ndiyaqonda ukuba ukuvuma ndithabathe inxaxheba kuthetha ukuba ndiyaavuma:

- Ndivumele umsebenzi wam obhaliwe ukuba ajongisiswe, xa ubumfihlo bugciniwe
- Mawaziswe ngeziphumo zophando.

Iinjongo zolu phononongo kukufezekisa isiDanga sePhD ekuFundiseni isiNgesi kwiziThethi Zezinye iiLwimi.

Ndiyaqonda ukuba ulwazi endilunikezile luza kuhlala luyimfihlo:

- Ukuthabatha kwam inxaxheba kungokuzithandela,
- Ndingakhetha ukungathathi inxaxheba kwinxalenye ethile okanye kuphononongo lonke, kwaye
- Ndingarhoxa nangaliphi inqanaba ngaphandle kokohlwaywa okanye ndicuthelwe amalungelo nangayiphi indlela.

Igama lomthabathi-nxaxheba

Utyikotyo

Umhla.....

Igama lomphandi

Inombolo yomfundi.....

Utyikityo.....

Umhla

Appendix B: Marking Grid (Content/Organisation – 25, Vocabulary, Language Usage, Mechanics – 25) ENG1501

Table: A1: Existing scale

Mark out of 25 for content/organisation:		
SCORE	LEVEL	CRITERIA
25-19 (100%-76%)	1 EXCELLENT TO VERY GOOD	<p>Content: focused on assigned topic, thoroughly developed, clearly demonstrating the skills required by the NQF criteria (e.g. familiarity with - recognising and recalling - the subject matter; understanding it; application of this information; analysis, for instance of relationships; evaluation, for example critiquing different approaches)</p> <p>Organisation: generating a piece of writing (such as an essay) with ideas clearly stated, succinct, well-organised, logically sequenced, cohesive, well-supported</p>
18-14 (75%-56%)	2 GOOD TO AVERAGE	<p>Content: fairly sound demonstration of skills, mostly relevant to topic, lacks detail</p> <p>Organisation: loosely organised, logical but incomplete sequencing and signposting</p>
13-8 (54%-32%)	3 FAIR TO SHAKY: AT RISK	<p>Content: not enough substance or relevance, insufficient support for ideas</p> <p>Organisation: ideas confused or disconnected, not enough logical sequencing or development, little signposting</p>
7-0 (30%-0%)	4 VERY SHAKY	<p>Content: not pertinent or not enough material to evaluate</p> <p>Organisation: does not communicate, no organisation or not enough material to evaluate</p>

Mark out of 25 for form (vocabulary, language usage, mechanics):		
SCORE	LEVEL	CRITERIA
25-19 (100%-76%)	1 EXCELLENT TO VERY GOOD	<p>Vocabulary: sophisticated range, effective word/idiom choice, mastery of word form, appropriate register</p> <p>Language usage: effective complex constructions, few language problems (agreement, tense, number, word order, articles, pronouns, prepositions)</p> <p>Mechanics: mastery of presentation: neatness, spelling, punctuation, capitalisation, paragraphing and essay structure; meticulous and consistent referencing of sources used</p>
18-14 (75%-56%)	2 GOOD TO AVERAGE	<p>Vocabulary: satisfactory range, occasional issues of word choice, idiom, form, usage, but meaning not obscured</p> <p>Language usage: effective simple constructions, minor problems in complex constructions, several language issues but meaning seldom obscured</p> <p>Mechanics: occasional problems in mechanics</p>

13-8 (54%- 32%)	3 FAIR TO SHAKY: AT RISK	Vocabulary: small range, frequent issues of word/idiom, choice, usage Language usage: major problems in simple/complex constructions, frequent language issues including sentence construction problems, meaning confused or obscured Mechanics: frequent problems with mechanics, untidy handwriting, meaning confused or obscured
7-0 (30%- 0%)	4 VERY SHAKY	Vocabulary: essentially translation from mother tongue, little knowledge of English vocabulary, idioms, word forms, or not enough material to evaluate Language usage: virtually no mastery of sentence construction, dominated by problems, does not communicate, or not enough material to evaluate Mechanics: no mastery of conventions, dominated by problems in mechanics, handwriting illegible, or not enough material to evaluate

Appendix C: MODULE FORM

1 Module title and code

Introduction to English Literary Studies: Code ENG1501

2 Module level

NQF Level 5

3 Credit attached to the module

12

4 Field and sub-field of the module

SAQA Field 4: Language, Literacy and Communication

5 Purpose of the module

This module forms part of first-year English studies when taken in conjunction with the module 'Introduction to Applied English Language Studies' (ENG1502).

This module will establish a literary and academic foundation for English studies. It will introduce students to representations of diversity in literature. Students credited with this module will be able to apply appropriate reading strategies to a wide variety of literary and non-literary texts in English. They will also be able to demonstrate basic skills of writing academic English.

6 Pre- and co-requisites

The following levels of learning ought to be in place to ensure successful completion of this unit standard:

The credit calculation is based on the assumption that students have successfully completed Grade 12 and are already competent in terms of the following:

- the ability to read texts in a focused and critical way;
- the ability to communicate information coherently and reliably in the language of tuition using basic conventions of academic discourse;
- the ability to take responsibility for their own learning in a distance learning environment.

7 Specific Outcomes and assessment criteria

A range of tasks in study guides, tutorial letters, assignments and examinations will show that students have achieved the following outcomes:

Specific Outcome 1:

Read a range of literary texts in different genres (poetry, prose and drama) with comprehension at an inferential level.

Assessment criteria:

A selection of literary texts is read and commented on, using acceptable academic discourse. Accepted conventions of academic discourse are applied.

Specific Outcome 2:

Demonstrate basic awareness of the creative choices made by writers of literary texts in English.

Assessment criteria:

- The dimensions of artistry and contrivance in the composition of literary texts in English are explored and explained through acceptable academic discourse.

Accepted conventions of literary criticism are applied.

8 Assessment strategy and plan

The student's mark will comprise a year mark that will be gained from one or more written assignments and a written examination of 2 hours at the end of the semester. First examiners will set and assess the assignments, tasks, activities and examination. In the case of examinations, second examiners will moderate questions, the marking process, and the marked scripts. Second examiners will also assist in the management of oral examinations. All examiners shall be senior academics or specialists in the field.

9 Syllabus

Selected literary texts;

Critical vocabulary pertaining to the interpretation of literature in English.

Prescribed texts:

Hard Times by Charles Dickens

Catcher in the Rye by John Salinger

The Road to Mecca by Athol Fugard

When Rainclouds Gather by Bessie Head

Seasons Come to Pass (Oxford University Press, eds. Moffett and Mphahlele)

Appendix D1: Email to Students

Dear Student

I am a doctoral student and long- serving tutor at Unisa, and am conducting research on assessment in the distance education context. My particular focus is on the validation of the rating scale for ENG1501 (please see information attached).

Your assignment has been randomly selected from the database to assist me in this research. Your name and student number will be deleted, so anonymity is assured. Please indicate if you would like to help me in this worthwhile project by signing one of the consent forms attached and returning it to me. An electronic signature is acceptable at present.

I hope that you can assist me and look forward to hearing from you soon.
All the best for your studies.

Regards
Maxine Ward-Cox
Student number: 2311194

Appendix D2: Email to Unisa Markers and Tutors

Dear xx

I am a PhD student and tutor at Unisa. The topic of my thesis is the validation of the rating scale for ENG1501, with the aim of either revising or replacing it. Could you assist me by providing feedback on the current marking grid attached? I would be interested in your input from the viewpoint of an e-tutor. I have obtained ethical clearance for the project.

If you are willing to assist with this research, please complete the attached questionnaire and return it and the signed consent form to me. An electronic signature will be sufficient if it is problematic to have the page scanned.

I am excited about this project and would like to think that the research results will benefit our students by improving formative and summative assessment.

I hope that you will participate in this project. Should you require any further information, please do not hesitate to contact me, either by email or telephonically. My contact details are provided below.

Many thanks for your attention to this matter. I look forward to hearing from you.

Regards

Maxine Ward-Cox

Appendix E: Assignment Memorandum

ENG1501 Assignment 1 Semester 2: 2016

Seasons Come to Pass

Instruction: Read the poem below, and then answer the questions that follow. Each question on the poem should be answered in paragraph form (10–15 lines) and written in full sentences. Remember to quote from the poem to substantiate your answers.

Small Passing (Ingrid de Kok)

1. The epigraph (the lines in italics directly below the title of the poem) introduces a stark contrast between the ‘small passing’ and the everyday suffering of black South Africans. By referring to the first section (lines 1–35) of the poem, explain how the poet creates this contrast.

The epigraph suggests that the loss of a white child is incomparable to the loss and suffering experienced daily by many black South African women. While this statement may seem harsh, it is not unwarranted. The poet creates a contrast between the passing of the child in the first stanza and the suffering of black people in the second stanza. The first stanza depicts the suffering of a woman who has given birth to a stillborn child, suggested by ‘the last push into shadow and silence’ (line 3). The speaker notes the now ‘useless wires and cords on your stomach’ (line 4), as well the futility of her breasts ‘still full of purpose’ (line 7). The woman who has lost her child searches for it in her house despite knowing it is not there, and recalls the doctor’s uncomfortable words that ‘It was just as well’ (line 11), and that she can always ‘have another’ (line 12). The first stanza ends with the provocative assertion that in ‘this country you may not / mourn small passings’ (lines 13–14). Despite this statement, the woman’s restlessness symbolises the genuine mourning of women of all backgrounds who experience the loss of a new-born child.

The second stanza of the poem offers the explanation of why her mourning is unwarranted in the South African context, marked by the word ‘See’ (line 15). The speaker says that a newspaper boy ‘in the rain / will sleep tonight in a doorway’ (lines 15–16), a woman in a ‘bus line’ will be displaced next month; a baby will be sent to live with its relatives, who are also tired and sickly. She adds that Mandela’s daughter is growing up without her father because he is imprisoned, and yet another woman ‘moves so slowly / as if in a funeral rite’ (lines 29–30) while dusting photographs of other children instead of spending time with her own. The losses of black South Africans depicted in stanza two seem far greater than the loss of one baby by a woman who can easily ‘have another’, whereas the people mentioned in the second stanza are offered no alternatives to their fate; the nannies in stanza three only have their ‘legal gatherings’ (line 32) to console them.

2. Identify the tone of the poem, and explain how it contributes to its meaning.

The tone in the first stanza is wretched and gloomy because of the woman’s loss, while also being accusatory because she is not ‘allowed’ to feel this way. The words ‘In this

country you may not suffer' (lines 1–2) and 'mourn' (lines 14–15) contribute to the harsh and biting tone. The second stanza provides the evidence to support the accusation of the first stanza, and the tone is an indication of this. As the poem progresses, the tone becomes darker as the reader is made aware of severe and ongoing difficulties faced by South Africans on a daily basis, such as the child who is 'shot running, / stones in his pocket' (lines 39–40), or another child whose stomach is full of 'hungry air' (line 42). In this manner, the tone contributes to the meaning of the poem by underscoring the loss of the white woman who gave birth to the stillborn child, but more importantly, the daily losses experienced by other women and children. In the third section of the poem there is a significant change in tone as the speaker tells of a group of mothers who 'will not tell you your suffering is white', who 'will not say it is just as well', and who will not 'compete for the ashes of infants' (lines 50–52). These women provide comfort to those who have lost their children, despite their race, and despite their own loss. The tone of this last section reflects the comfort, warmth and support provided by these women who 'stroke your flat empty belly, / let you weep with us in the dark, / and arm you with one of our babies/ to carry home on your back' (lines 55–58).

3. This poem is an example of free verse, which means that it has no set structure. However, the poet uses a number of poetic devices to create rhythm and form. Identify the main sound device the poet employs, and discuss its effect by referring to at least three examples from the poem.

The most prominent sound device employed in this poem that contributes to its form and rhythm is alliteration. In the first stanza, the poet repeatedly uses s-sounds to emphasise loss, as depicted in 'suffer the death of your stillborn, / remember the last push into shadow and silence, / the useless wires and cords on your stomach, / the nurse's face, the walls, the afterbirth in a basin' (lines 2–5).

The repetition of s-sounds here underscores the suffering and the silence experienced by the woman who has lost her child, while also highlighting that she herself should remain silent, and not complain about it.

In the second stanza there is frequent repetition of b-sounds, as can be seen in the lines 'The baby in the backyard now / will be sent to a tired aunt, / grow chubby, then lean, / return a stranger' (lines 20–23). This repetition emphasises the loss of babies other than the woman's in the first stanza of the poem, and highlights the 'perspective' that the woman should have about her own loss when compared to others' losses. So too the alliteration in stanza four, 'Girls carrying babies / not much smaller than themselves' (lines 43–44) seeks to suggest this perspective – surely these losses are far greater than the loss of one?

4. Quote two similes from the poem, and explain what they mean. Also consider how these similes develop meaning in the poem as a whole.

There are two important similes to be found in section one of the poem. The first is: 'Clumsy woman, she moves so slowly / as if in a funeral rite' (lines 29–30). The woman, who drops photographs of 'other children' while dusting (line 27), moves very slowly, as if she is part of a funeral ceremony or ritual. This suggests that the woman is mourning not being able to spend time with her own children as a result of her having to clean someone else's home. This simile helps to develop the theme of the suffering of others that underlies the poem.

In stanza three, the speaker tells of the nannies and how ‘They talk about everything, about home, / while the children play among them, / their skins like litmus, their bonnets clean’ (lines 33–35). The skin of these children, who are being looked after by the nannies, is compared to ‘litmus’. Litmus can refer to the chemical used to determine the acidity or alkalinity of a solution, but it can also refer to an important or revealing test. In the context of this poem, the second denotation of the word is more likely to be implied here. The children’s skin is compared to a litmus test as it reveals the distinctive racial divide between the implied worker class black nannies and the carefree white children. The white children playing with their ‘bonnets clean’ reveal the distinct separation between races in South Africa and, by implication, the suffering of those who take care of the children of others and can only ‘talk’ about home.

5. Carefully consider lines 53–58, and explain how they contradict the rest of the poem.

From line 1–52, the speaker seems to suggest that the epigraph is correct in its implication that the loss of one small white child cannot possibly compare to the loss of so many black lives, both physically and emotionally. The first sections of the poem suggest that the woman who has lost her child should not mourn this loss, as there are far greater losses in the country that are not mourned.

However, the final lines of the poem suggest that mothers, regardless of race or creed, all know and understand loss. More specifically, ‘these mothers’ (line 46), who are also nannies taking care of white children, and who suffer losses daily, would not assert that the life of any child does not matter. The implication here is that, despite the fact that these mothers cannot see their own children as a result of class divides and racial segregation in the country, they would welcome all mothers who have suffered the loss of a child.

Unlike the man in the epigraph to the poem who tells the mother to ‘stop mourning’, these mothers will allow the woman to ‘weep with us in the dark’ (line 6) and to become one of them. The final lines of the poem thus imply the intimacy and mutual understanding that overcome all notions of racial and class segregation since experiences of loss and suffering are universal.

Appendix F: Rasch Analysis: Existing Rating Scale

Reliability

Table A.2: Summary of 60 measured Persons: Markers 1 – 6, Scripts 1 – 60

M1_6_60_T.csv	M1_6_60_T_results.txt
INPUT: 60 PERSON 7 ITEM REPORTED: 60 PERSON 7 ITEM 84 CATS WINSTEPS 4.1.0	
SUMMARY OF 60 MEASURED PERSON	
REAL RMSE .06 TRUE SD .19 SEPARATION 3.09 PERSON RELIABILITY .91	
MODEL RMSE .05 TRUE SD .19 SEPARATION 3.86 PERSON RELIABILITY .94	
S.E. OF PERSON MEAN = .03	
PERSON RAW SCORE-TO-MEASURE CORRELATION = .99	
CRONBACH ALPHA (KR-20) PERSON RAW SCORE "TEST" RELIABILITY = .92 SEM = 22.88	

	TOTAL SCORE	COUNT	MODEL MEASURE	S.E.	INFIT MNSQ	OUTFIT ZSTD	MNSQ	ZSTD
MEAN	369.3	8.0	-.09	.08	.99	-.3	.96	-.3
P.SD	76.9	.0	.45	.00	.80	1.5	.75	1.4
S.SD	78.3	.0	.46	.01	.81	1.5	.77	1.4
MAX.	550.0	8.0	1.09	.09	3.28	3.0	3.07	2.8
MIN.	194.0	8.0	-1.00	.07	.16	-2.7	.17	-2.6

Table: A.3: Summary of 8 measured Items: Markers 1 – 8, Scripts 1- 60

SUMMARY OF 8 MEASURED ITEM									
	TOTAL SCORE	COUNT	MODEL MEASURE	S.E.	INFIT MNSQ	OUTFIT ZSTD	MNSQ	ZSTD	
MEAN	1384.9	30.0	.00	.04	.96	-.9	.96	-.9	
P.SD	93.3	.0	.14	.00	.69	3.0	.70	3.0	
S.SD	99.7	.0	.15	.00	.74	3.2	.75	3.2	
REAL RMSE .04 TRUE SD .14 SEPARATION 3.09 ITEM RELIABILITY .91									
MODEL RMSE .04 TRUE SD .14 SEPARATION 3.54 ITEM RELIABILITY .93									
S.E. OF ITEM MEAN = .05									

ITEM RAW SCORE-TO-MEASURE CORRELATION = -1.00

Global statistics: please see Table 44.

UMEAN=.0000 USCALE=1.0000

Content (C) and Language (L), Markers 2-6

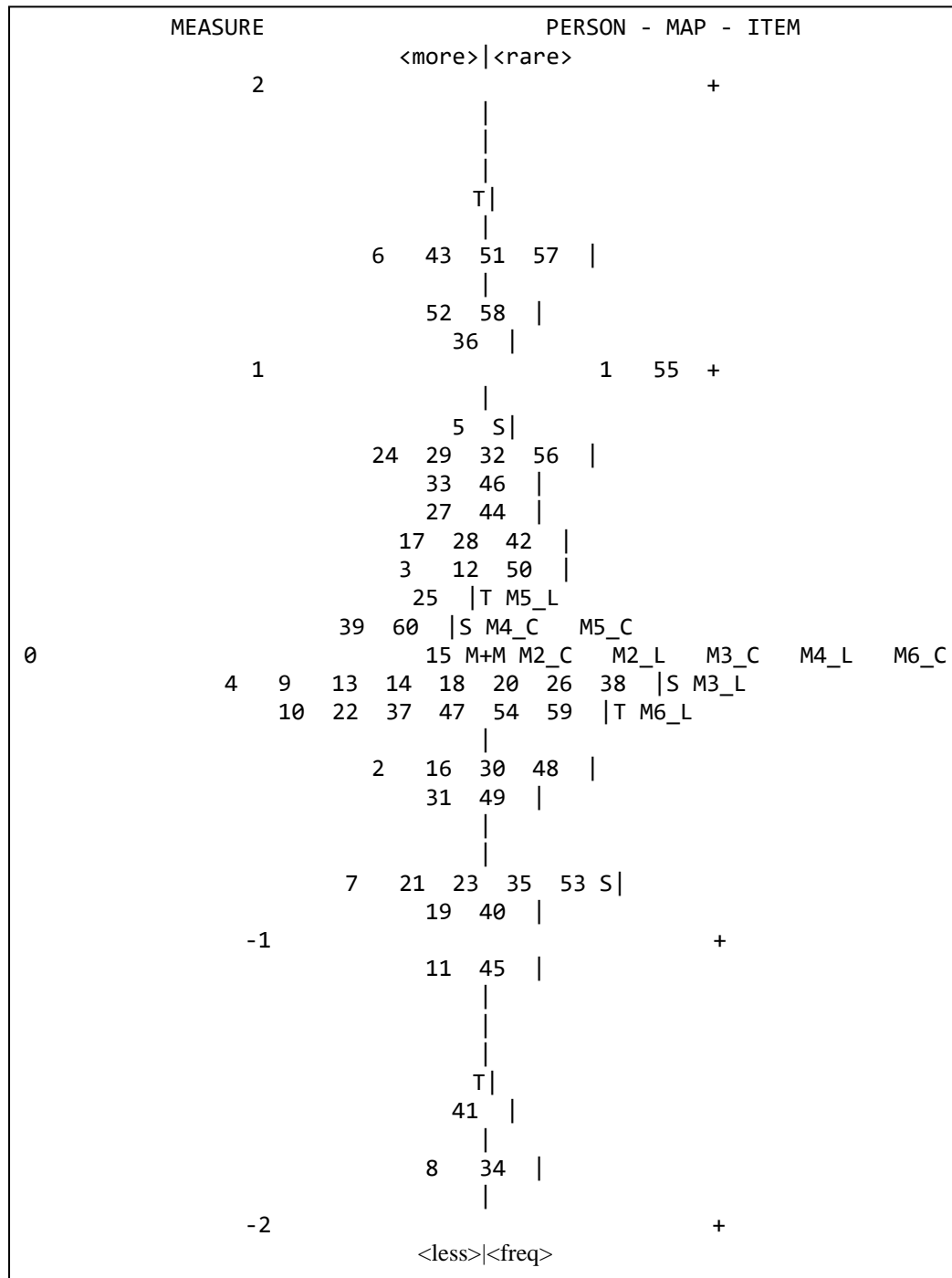


Figure A.1: Summary of content and language: Markers 2 – 6, scripts 1 - 60

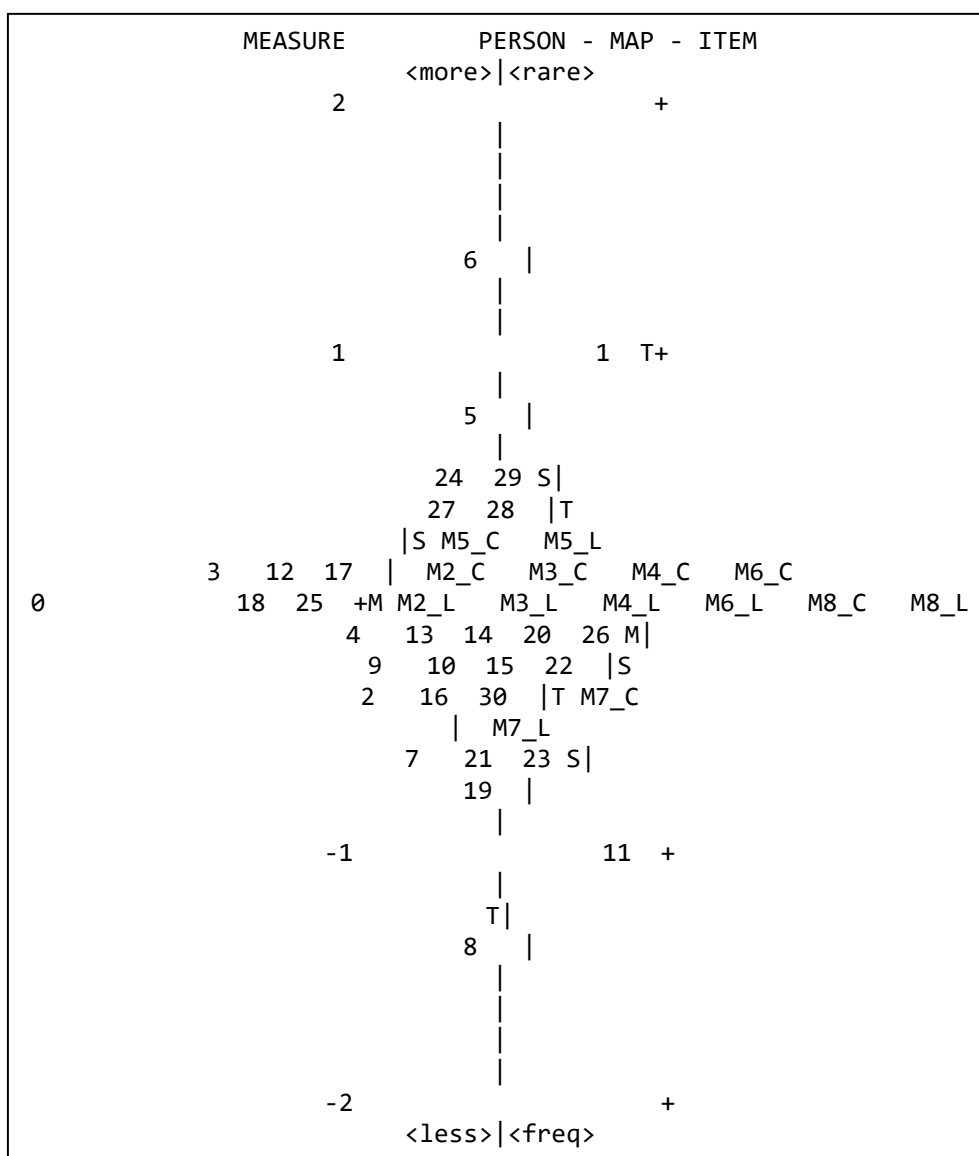


Figure A.2: Summary of content and language: Markers 2 – 8, scripts 1 - 30

Appendix G: Extracts from correspondence with panel members

Extract 1

“I’ll just write a few ideas more or less ‘as the spirit leads’. Let me start with the general subject of the well-established practice of rating scales for written work. While I consider these invaluable for young people without much experience, not least because they ‘force’ one to decide what the descriptions in the categories would look like in practice, as well as to consider and compare each script both to the grid and to the work of other students, I think that experienced markers probably need to do no more than keep rating scales in mind in broad terms. This these markers would do while capably turning out sound and consistent evaluations almost automatically, though that word might not strike quite the right note, since their marking is not a fluke but a carefully developed skill honed over many years and also does not happen effortlessly, but with constant concentration and application! Rating scales are a good touchstone for anyone for reference at the start of a new assignment or exam or group of students but I believe that constant recourse to them is not needed if one is experienced. What the criteria would be for someone to fall into that category of marker, and who would do the categorising I can’t say, though.

Of course, one of the great shortcomings of rating scales of the type we use, as pointed out by some of your respondents also, is that discrepancies can arise because of form and meaning being interdependent, to the detriment of the latter. It often appears as though very poor language and expression constrain expression to the extent that a student cannot articulate what he or she seems to understand about a text. A particular style of language use such as pomposity or excessive informality, chattiness or stiffness or the terminology of a seemingly all-encompassing ideology, has a comparable deleterious effect on the presentation of ‘content’. In general, I am in favour of trying to acknowledge the spark of insight cautiously, and explaining to the style bandits briefly what the problem is, how it affects their work, and how to do something about it. Whether this has ever been any good to them I do not know!

I agree with your respondent(s) that the rating scales do not provide for plagiarism, but that there is a need to try to acknowledge it because the practice is ubiquitous now and nothing seems to deter students from either straight copying or judicious paraphrase without acknowledgment. However I cannot think of a way to insert all the possible variations into the current content scale, and would propose penalising the guilty on the overall total mark (with full explanation of the reasons and what the offense is, including naming the source or sources) along the lines of naught for straight copying of more than half of a piece of work, minus 50% for paraphrasing more than half, with the same condition as above, and corresponding lesser percentages off for lesser offenses. I think the total mark should be the basis, since the structure of argument, as well as language and expression are also either rendered in facsimile, or copied with similar syntax and synonyms so that there isn’t an accurate record of a student’s own writing and language skills in plagiarised work.

As far as inaccurate or lack of referencing go, to my mind this is a peripheral issue that merely needs to be pointed out to students, with brief advice on how to correct the problem and more on where to find full information to help them do so.

With reference to the scales themselves, I agree totally with the person who suggested that the categories are too broad, particularly the top one, where ‘excellent’ needs to be so much more than just meeting all expectations, and the one where the pass mark falls. If there is a clear description of what is ‘passable’ ... marking would be much easier for the inexperienced particularly. With regard to content, a pass would require some attempt at relevance that is more than a paraphrase or summary of a text; a minimum of sentences in paragraphs and some logical organisation for an essay, and a sufficient quantity (not just a paragraph or two). For language I suggest the key criteria for passing would be intelligibility (in spite of frequent language errors and problematic expression as well as lack of organisation) and some indication of a minimum awareness of appropriate register and vocabulary.

Lastly, and a purely personal comment, I myself never consult the ‘form’ table, I suppose because of having mentally fused the two somewhere in the more than forty years I have been marking. I would just hate to give two separate marks and then add them up, to me holism is the essence of the process”.

Extract 2

“I also like the idea of drawings, but I am not sure how one escapes these being culture-loaded nowadays. Does a ‘skedonk’[old, dilapidated car] represent poverty or the first step upwards to a rural student who has never been in a car? How could one be sure that the ‘sign’ represented the truth any better than words did? I suggest that you abandon the ideas of pictures, especially as it is really worthwhile to have the grid on one page.

Maybe a covering letter encouraging lecturers to convey the info in a visual form to their students would work in a non-correspondence institution, but if your aim is to design a Unisa grid that probably wouldn’t work either. One could maybe include a note suggesting the student visualise their work in terms of a drawing with the study materials – then at least one has tried to provide a concrete image”.

Extract 3 (reference to Model 2 grid)

“Yes, I think the student should get feedback via a copy of the grid, but my experience is that they receive this info on cell phones, don’t print it and so it doesn’t register. If they received one at the beginning of the course, all lecturers need to do is to say D4 because/ despite The lecturer can highlight the deciding factor or exclude any factor which the student did not display. Saves paper, and saves the lecturer having to think of the words to use – the grid provides the vocab for content, though they might actually have to create the comments on language”.

Appendix H : Statistical information: Model 1

Table A.5: Summary of 60 measured Persons: Markers 1 – 6, Scripts 1 – 60

Modell1_59scripts_5.csv Modell1_59scripts_5_res.txt May 27 2019 8:33									
INPUT: 60 PERSON 5 ITEM REPORTED: 60 PERSON 5 ITEM 71 CATS WINSTEPS 4.1.0									
SUMMARY OF 60 MEASURED PERSON									
	TOTAL SCORE	COUNT	MODEL MEASURE	S.E.	INFIT MNSQ	OUTFIT ZSTD	MNSQ	ZSTD	
MEAN	210.4	5.0	.11	.10	.95	-.4	.98	-.4	
P.SD	64.0	.1	.63	.02	1.09	1.6	1.17	1.6	
S.SD	64.5	.1	.63	.02	1.10	1.6	1.18	1.6	
MAX.	350.0	5.0	1.25	.23	4.83	3.7	4.91	3.8	
MIN.	42.0	4.0	-2.34	.08	.00	-3.5	.00	-3.4	
REAL RMSE .12 TRUE SD .62 SEPARATION 5.30 PERSON RELIABILITY .97 MODEL RMSE .10 TRUE SD .62 SEPARATION 6.15 PERSON RELIABILITY .97 S.E. OF PERSON MEAN = .08									
PERSON RAW SCORE-TO-MEASURE CORRELATION = .98 CRONBACH ALPHA (KR-20) PERSON RAW SCORE "TEST" RELIABILITY = .97 SEM = 10.90									

Table A.6: Summary of 6 measured Items: Markers 1 – 6, scripts 1 - 60

SUMMARY OF 5 MEASURED ITEM									
	TOTAL SCORE	COUNT	MODEL MEASURE	S.E.	INFIT MNSQ	OUTFIT ZSTD	MNSQ	ZSTD	
MEAN	2525.4	59.8	.00	.03	.99	-.5	.98	-.6	
P.SD	159.3	.4	.12	.00	.53	2.9	.52	2.9	
S.SD	178.1	.4	.14	.00	.60	3.2	.58	3.2	
MAX.	2753.0	60.0	.21	.03	1.88	3.8	1.86	3.8	
MIN.	2264.0	59.0	-.17	.03	.38	-4.3	.38	-4.4	
REAL RMSE .03 TRUE SD .12 SEPARATION 3.90 ITEM RELIABILITY .94 MODEL RMSE .03 TRUE SD .12 SEPARATION 4.33 ITEM									

RELIABILITY .95

S.E. OF ITEM MEAN = .06

REAL RMSE .03 TRUE SD .12 SEPARATION 3.90 ITEM
RELIABILITY .94

MODEL RMSE .03 TRUE SD .12 SEPARATION 4.33 ITEM
RELIABILITY .95

S.E. OF ITEM MEAN = .06

ITEM RAW SCORE-TO-MEASURE CORRELATION = -1.00

Global statistics: please see Table 44.

UMEAN=.0000 USCALE=1.0000

Appendix I : Statistical Results of Model 2

Table A.7: Summary of 60 measured Persons: Markers 1 – 6, Scripts 1 – 60

PERSON RAW SCORE-TO-MEASURE CORRELATION = .99									
CRONBACH ALPHA (KR-20) PERSON RAW SCORE "TEST" RELIABILITY = .99 SEM = 8.71									
TABLE 3.1 Data new scale trial Model 2 Model2_60_results.txt Jan 17 2019 12:39									
INPUT: 60 PERSON 5 ITEM REPORTED: 60 PERSON 5 ITEM 73 CATS WINSTEPS 4.3.0									
SUMMARY OF 60 MEASURED PERSON									
	TOTAL SCORE	COUNT	MODEL MEASURE	S . E .	INFIT MNSQ	OUTFIT ZSTD	MNSQ	ZSTD	
MEAN	212.4	5.0	.14	.13	.86	-.32	.87	-.30	
SEM	9.3	.0	.17	.00	.10	.16	.10	.16	
P.SD	71.2	.0	1.27	.04	.79	1.19	.80	1.19	
S.SD	71.8	.0	1.28	.04	.80	1.20	.80	1.20	
MAX.	367.0	5.0	3.16	.23	4.30	3.11	4.43	3.18	
MIN.	32.0	5.0	-2.89	.08	.03	-2.52	.02	-2.52	
REAL RMSE .15 TRUE SD 1.26 SEPARATION 8.21 PERSON RELIABILITY .99									
MODEL RMSE .14 TRUE SD 1.26 SEPARATION 9.02 PERSON RELIABILITY .99									
S.E. OF PERSON MEAN = .17									
PERSON RAW SCORE-TO-MEASURE CORRELATION = .99									
CRONBACH ALPHA (KR-20) PERSON RAW SCORE "TEST" RELIABILITY = .99 SEM = 8.71									

Table A.8: Summary of 6 measured Items: Markers 1 – 6, scripts 1 - 60

SUMMARY OF 5 MEASURED ITEM									
	TOTAL SCORE	COUNT	MODEL MEASURE	S . E .	INFIT MNSQ	OUTFIT ZSTD	MNSQ	ZSTD	
MEAN	2548.8	60.0	.00	.04	.91	-.91	.87	-.97	
SEM	28.2	.0	.04	.00	.22	1.30	.20	1.10	
P.SD	56.4	.0	.07	.00	.45	2.60	.40	2.20	
S.SD	63.0	.0	.08	.00	.50	2.91	.45	2.46	
MAX.	2617.0	60.0	.13	.04	1.47	2.24	1.42	1.93	

MIN.	2445.0	60.0	-.09	.04	.42	-4.00	.42	-3.71	
REAL RMSE .04 TRUE SD .06 SEPARATION 1.55 ITEM RELIABILITY .71									
MODEL RMSE .04 TRUE SD .06 SEPARATION 1.73 ITEM RELIABILITY .75									
S.E. OF ITEM MEAN = .04									
ITEM RAW SCORE-TO-MEASURE CORRELATION = -1.00									
Global statistics: please see Table 44.									
UMEAN=.0000 USCALE=1.0000									

Appendix J: Final Grid Model 1

Classification →	Exceptional (Distinction)	Excellent (Distinction)	Good to Above Average	Average	'Borderline' FAIL	Seriously at risk: Fail
Mark →	25-22	22-19	18-15	14-12.5	11.5-9	8-0
Content/organisation. Criteria a. Insight: To what extent does the answer show maturity, understanding and originality? b. Awareness of stylistic/ technical features: Are these accurately demonstrated? c. Substantiation: Is the answer supported by appropriate reference to the text? d. Relevance: To what extent has all relevant information been included? Has the question been fully answered? e. Coherence – Does the answer flow together? NB MARK GLOBALLY	SUMMARY Exceptional insight and organisation. a. Original, sensitive, mature interpretation and insight. b. Exceptional, original and sensitive c. Unfailingly well-supported, shows depth and insight d. Extremely relevant, well chosen, valid ideas, all points fully covered. e. Shows exceptional focus, cohesion, seamless organisation	SUMMARY: Mature, original, comprehensive. a. Thorough, incisive, original b. Excellent, good examples. c. Extremely well supported with apt examples. d. Extremely relevant, all points covered. e. Exceptionally well structured, focused, coherent.	SUMMARY: Sufficient understanding well organised, comprehensive relevant a. Good grasp of issues, some originality b. Well-demonstrated appreciation. c. Generally well substantiated. d. Mostly relevant, most issues addressed. e. Well organised, and coherent.	SUMMARY: Adequate understanding lacks originality and depth. a. Adequate, lacks depth, little originality b. Features occasionally discussed, usually correct c. Some substantiation, but mainly just thoughts on the question. d. Fairly relevant, point sometimes missed. e. Loosely organised but still coherent.	SUMMARY, Disconnected, largely irrelevant. a. Insight inadequate, little understanding of issues b. Features seldom discussed, shows lack of knowledge c. Not enough substance or relevance, insufficient support for ideas d. Many statements lack relevance e. Ideas confused or disconnected, little logical sequencing or development.	SUMMARY: serious errors, irrelevant, confused. plagiarised. a. Serious errors of understanding, extremely little evidence of knowledge of text. b. Features ignored. c. Unsubstantiated. d. Irrelevant 'misses the point'. e. Incoherent., disjointed. Plagiarised OR Not enough to evaluate.
Sub-total 25 marks						
Criteria Language and style a. Vocabulary- is there a sufficient range of vocabulary and effective word choice? b. Register and tone- appropriate for academic writing or too informal or pretentious? c. Language errors – are errors negligible or are they frequent, <u>impeding meaning?</u>	SUMMARY: Clear, fluent, articulate, sophisticated a. Excellent range and word choice. b. Very appropriate. Clear, formal c. Negligible or no language errors:	SUMMARY: Clear, fluent, articulate a. Good range of vocabulary; very appropriate word choice b. Appropriate register, competently used c. Occasional language errors but <u>meaning not impeded or confused</u>	SUMMARY: Clear despite errors a. Word choice and range generally sufficient b. Generally appropriate. c. Some language errors, <u>but meaning not impeded.</u>	SUMMARY: Adequate but pedestrian (dull) a. Small but adequate range, some errors of word choice. b. Occasional lapses c. Frequent language errors, but <u>meaning seldom impeded.</u>	SUMMARY Meaning seriously impeded a. Very small range frequent errors forward choice. b. Inappropriate (too informal or too verbose). c. Frequent and serious errors, <u>meaning confused or obscured</u>	SUMMARY: Meaning severely impeded due to frequent and fundamental errors a – c Serious and distracting errors, frequently barely intelligible. OR Not enough to evaluate. OR . Plagiarised

Appendix K : Proposed Marking Grid ENG1501 Literary Assignments

Content and organization NB Mark globally	A. Outstanding HIGH DISTINCTION	B. Excellent to very good DISTINCTION	C. Good to fair	D. Adequate	E. Fail AT RISK	F. Fail SERIOUSLY AT RISK
	SUMMARY	SUMMARY	SUMMARY	SUMMARY	SUMMARY	SUMMARY
<div> <div></div> <div>Language and Style NB Mark globally</div> </div>	Mature, original, comprehensive, very logical, exceptional insight and organisation.	Excellent understanding and organisation, comprehensive and relevant	Sufficient understanding, well organised, comprehensive and relevant	Adequate understanding, lacks originality and depth, misses some important points	Misses the point, disconnected, largely irrelevant, some evidence of plagiarism	Serious errors, irrelevant, confused. Largely plagiarised; not enough to evaluate.
	<p>a. Insight: Original, sensitive, mature interpretation and insight.</p> <p>b. Awareness of stylistic/technical features: Exceptional, original and sensitive</p> <p>c. Unfailingly well-supported, shows depth and insight</p> <p>d. Extremely relevant, thought-provoking</p> <p>e. Structure shows exceptional focus, cohesion, seamless organisation</p>	<p>a. Insight: Thorough, incisive, original</p> <p>b. Awareness of stylistic/technical features: Excellent, good examples.</p> <p>c. Extremely well supported with apt examples.</p> <p>d. Extremely relevant, all points covered.</p> <p>e. Extremely well structured, focused, coherent.</p>	<p>a. Insight: Good grasp of issues, some originality.</p> <p>b. Awareness of stylistic/technical features: Well-demonstrated appreciation.</p> <p>c. Well substantiated.</p> <p>d. Mostly relevant, most issues addressed, points of question covered.</p> <p>e. Well organised.</p>	<p>a. Insight: Adequate, lacks depth, little originality</p> <p>b. Awareness of stylistic/technical features: Usually correct, but insufficiently discussed</p> <p>c. Some substantiation, but mainly just thoughts on the question.</p> <p>d. Fairly relevant, point sometimes missed.</p> <p>e. Loosely organised but still coherent.</p>	<p>a. Insight inadequate, little understanding of issues.</p> <p>b. Awareness of stylistic/technical features: Features seldom/inadequately discussed.</p> <p>c. Not enough substance or relevance, insufficient support for ideas</p> <p>d. Many irrelevant statements</p> <p>e. Ideas confused or disconnected, not enough logical sequencing, little signposting</p> <p>OR</p> <p>f. Evidence of plagiarism (whole sentences/paragraphs)</p>	<p>a. Insight: Serious errors.</p> <p>b. Awareness of stylistic/technical features: Features ignored.</p> <p>c. Mostly unsubstantiated.</p> <p>d. Irrelevant 'misses the point'.</p> <p>e. Incoherent., disjointed.</p> <p>OR</p> <p>f. Largely plagiarised</p> <p>g. not enough to evaluate.</p>
1. Outstanding. Vocabulary: apt, sophisticated; Correct formal register effectively used; Very few language problems; <u>Meaning not impeded.</u>	A1 100%-85%	B1 84% – 75%	C1 74%-70%	D1 69%-65%		
2. Excellent to very good. Good use of vocabulary; Correct register; occasional language errors but <u>meaning not impeded or confused</u>	A2 84% – 75%	B2 74%-70%	C2 69%-65%	D2 64%-60%		
3. Good Satisfactory vocabulary. Some errors of word choice and register but <u>meaning seldom obscured.</u>		B3 69%-65%	C3 64%-60%	D3 59%-56%		
4. Adequate Small range of vocabulary; Frequent problems with register, language, word choice, sentence structure and mechanics; <u>Meaning sometimes obscured or confused</u>			C4 59%-56%	D4 55%-50%	E4 49%-40%	F4 39%-30%
5. FAIL: AT RISK Little knowledge of English vocabulary; poor register ;Numerous language problems <u>that seriously impede communication</u> ; Not enough to evaluate				D5 49%-40%	E5 39%-30%	E6 29%-25%
6. FAIL: SERIOUSLY AT RISK Numerous problems (register, word choice, sentence structure, and mechanics) <u>that seriously impede communication</u> ; Not enough to evaluate OR plagiarised				D6 39%-30%	E6 29% -25%	F6 24%-0%

Appendix L: Statement of originality of topic

From: Malan, Dawie [mailto:Malandj@unisa.ac.za]

Sent: 06 November 2019 03:23 PM

To: Maxine Ward-Cox

Cc: Spencer, Brenda

Subject: RE: For your information: submission of M&D degrees to UIR - procedures

hello maxine, and brenda

no research projects locally, nor internationally, stand in the way of the originality of the title, ***Validation of a rating scale for distance education university student essays in a literature based module***

attached 1986 dissertation is for interest only, as the only one matching validation/validity + rating scale +student essays. well done maxine, well done brenda

dawie malan

librarian: human sciences

unisa library 7-17

pretoria 0003

malandj@unisa.ac.za

phone 27-12-4293212

Appendix M: Researcher's Curriculum Vitae (abbreviated)

Curriculum Vitae- Maxine Welland Ward- Cox

Personal Details

Full Name: Maxine Welland Ward- Cox

Address: 27 Disa Road

Tableview

Cape

7441

Telephone: (021) 5573314

Mobile: 083 2855971

E-mail: maxibob@telkomsa.net

Date of birth: 21 October 1949

Gender: Female

Nationality: South African

Education

1. Post-Graduate:

MA (TESOL) -passed *cum laude*

Date and institution 2013 Unisa

Supervisor: Prof. Brenda Spencer

BA (Hons) (TESOL) –passed *cum laude*

Date and institution: 2003 Unisa

Diploma for Special Education (School Libraries)

Date and institution: 1982- Unisa

BA (Hons) – English Literature

Date and institution 1976-Unisa

Secondary Teacher's Diploma

Date and institution: 1970- University of Cape Town

2. Bachelor of Arts:

1969 University of the Witwatersrand: Major subjects: English III, French III, Afrikaans en Nederlands II

Other Qualifications

November 2005- Accepted by High Court as a Sworn Translator – Afrikaans/English.

Employment History (abbreviated)

1971- 1982: Employed by WCED as a teacher (Post level 1) at various High Schools in the Cape Peninsula

Subjects taught: English first and second language (Grades 8-12): Afrikaans second language (Grade 8-11): History (Grade 8-12).

1982- 2004: Lecturer and then Senior lecturer of English and Communication at the College of Cape Town

Subjects taught: English, Afrikaans, Communication.

Main Duties: Co-ordination of Language Programmes, editing of various college publications.

Other achievements: Articles published in educational publications

2004- present:

Part-time tutor at Unisa Learning Centre

Subjects taught: English Literature (ENN101-D, ENN102-E), Training in Thinking Skills (TSK) Business Communication, Language and Learning Skills (LSK0108).

July –September 2005:

Part-time tutor at Damelin College

Subjects taught: LSK (Learning Skills), Business Communication (Unisa courses).

September 2005

Appointed as on-site tutor (English and Afrikaans) at International Colleges Group, Cape Town.

January 2006 – December 2006

Academic Administrator (ICG)

June 2006- present

1. Freelance author for Nasou Via Afrika.

Texts include:

Language for FET colleges NQF Level 3 and 4

Co-complier of *The Storyteller*, an anthology of short stories for Grade 12

2. Freelance tutoring and moderating at CPUT

Other Positions

1991, 2016: National examiner of English Olympiad (4000 candidates)

1992: Selected as Director of Literary Tour of England for prizewinners of English Olympiad

1990-1992: National examiner of Communication (N1) (3000 – 4000 candidates per examination).

Appendix N: Turnitin similarity index showing single source similarities (5% and above) excluding referenced quotations.

Turnitin Originality Report

- Processed on: 27-Jan-2020 17:23 SAST
- ID: 1247099251
- Word Count: 98785
- Submitted: 1

Validation of a rating scale for distance education university student essays in a literature-based module... *By Maxine Welland Ward-Cox*

Similarity Index

0%

Similarity by Source

Internet Sources:

0%

Publications:

0%

Student Papers:

0%

exclude quotations [include bibliography](#) [excluding matches < 5%](#) mode:

show highest matches together



Change mode [print](#) [download](#)

There are no matching sources for this report.